UNIVERSITÀ COMMERCIALE LUIGI BOCCONI
ISTITUTO METODI QUANTITATIVI

Bayesian variable and model selection for Customer Lifetime Value

Phd in Statistics (XIX Ciclo)

Director: Prof. Pietro Muliere

Candidate: Silvia Figini (935460DT)

Accademic Year 2005/06

Dubium sapientiae initium.

Cartesio

Contents

List of Figures

Acknowledgements

This thesis marks the achievement of my Ph.D. studies.

First of all, the bulk of my thanks goes to my dissertation advisor Prof. Paolo Giudici. Not only he has inspired me the topic of this dissertation, but also he drove me patiently in the development of the work with many valuable suggestions. I owe him a number of fruitful and stimulating advices of the greatest moments for this dissertation.

I would like to thank the Phd Director, Prof. Pietro Muliere to have provided me with statistical guidance and suggestions, as well as endless support and encouragement throughout my Phd years.

I am also grateful to Prof. Steve Brooks for the time and the attention he devoted to me during my visiting period at the University of Cambridge.

Thesis structure

This dissertation starts with a review on variable selection techniques and is composed of 6 chapters.
In Chapter 1 there is a review of the literature on variable selection and a new proposal for a general framework to improve the dimensionality reduction with attention on the relationships of Max-dependency, Max-relevance and Min-redundancy. The main novel contribution of this Chapter is a general formalisation for the variable selection problem based on a Lemma.

Chapter 2 starts with a new mathematical formalisation to define Customer Lifetime Value. Later I introduce the classical statistical tenure model based on Cox regression and I show the weaknesses and criticisms of such classical churn models. Then I point attention to new lifetime value models, in particular focusing on survival analysis in the point processes framework. I also put in this section a review on model adequacy and model comparison.

In Chapter 3 I propose a new Bayesian extension of lifetime value models, based on Bayesian methods to choose the most important variables and I present a novel Bayesian model based on point processes framework. I improve this model following some approach in literature based on Bayesian Model Averaging. Comments on the possible evolution of this approach with novel methodological contribution is then discussed, especially with reference to stratification and multi-level models. Finally, model search and model comparison for Bayesian lifetime value models is discussed.

Chapter 4 improves the previous results with the presentation of Bayesian stratified models, with fixed and random effects, and a discussion on the efficiency of partial likelihood methods. A new field of research (theoretical and applied) based on penalised likelihood methods, is then proposed, with a Bayesian extension and a discussion on computational issues. Estimation using shared frailty models can be performed with penalized likelihood methods and therefore in a simple and realistic Bayesian framework. Such a correspondence has been found for gamma and gaussian frailty models. In this Chapter we remark new computational contributions.

In Chapter 5 new methods for model selection are proposed. We remark that the first part of this Chapter is justified by an empirical analysis based on the application results; this Chapter is critical from an applicative and theoretical point of view. The aim is only to introduce new criteria to obtain clusters of models. We will improve this starting point of research. Starting from 5.2, we would like remark novel contribution in terms of model assessment and predictive performance, with particular attention to economic assessment and decision making, based on loss curves.

Chapter 6 shows the conclusions and future lines of research.

Chapter 1

Variable Selection

1.1 The variable selection problem: introduction

The problem of variable selection, or as in the most recently literature, feature selection, is one of the most pervasive problems in statistical applications. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use.

Suppose Y a variable of interest, and $X_1, ..., X_p$ a set of potential explanatory variables or predictors, are vectors of n observations. The problem of variable selection, or subset selection as it is often called, arises when one wants to model the relationship between Y and a subset of $X_1, ..., X_p$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when p is large and $X_1, ..., X_p$ is thought to contain many redundant or irrelevant variables.

The fundamental developments in variable selection seem to have occurred either directly in the context of the linear model or in the context of general model selection. Historically, the focus began with the linear model in the 1960s when the first wave of important developments occurred and computing was expensive. The focus on the linear model still continues, in part because it is analytic tractability greatly facilitates insights, but also because many problems of interest can be posed as linear variable selection problems. For example, for the problem of nonparametric function estimation, Y represents the values of the unknown function, and $X_1, ..., X_p$ represent a linear basis such as a wavelet basis or a spline basis. However, as advances in computing technology have allowed for the implementation of richer classes

of models, treatments of the variable selection problem by general model selection approaches are becoming more prevalent.

One of the fascinating aspects of the variable selection problem is the wide variety of methods that have been brought to bear on the problem. Because of space limitations, it will of course be impossible to even mention them all, and so I have only focused on a few to illustrate the general thrusts of developments. An excellent and comprehensive treatment of variable selection methods prior to 1990 can be found in Miller (1990). As we will see, many promising new approaches have appeared over the last decade.

A distinguishing feature of variable selection problems is their enormous size. Even with moderate values of p, computing characteristics for all 2^p models is prohibitively expensive and some reduction of the model space is needed.

Once attention is reduced to a manageable set of models, criteria are needed to select a subset model. The earliest developments of such selection criteria, again in the linear model context, were based on attempts to minimize the mean square error of prediction. Different criteria corresponded to different assumptions about which predictor values to use, and whether they were fixed or random, see Hocking (1976) and the references therein. Many interesting criteria have been proposed in the literature, see for example Shao (1996), Clyde and George (2000), George and Foster (2000). An alternative is selection based on predictive error estimates obtained by intensive computing methods such as the bootstrap and cross validation e.g. Shao (1996).

1.2 A review on variable selection methods

There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance. Some methods put more emphasis on one aspect than another. Here we focus mainly on constructing and selecting subsets of features that are useful to build good predictors. This contrasts with the problem of finding or ranking all potentially relevant variables. Selecting the most relevant variables is usually suboptimal for building a predictive model, particularly if the

variables are redundant. Conversely, a subset of useful variables may exclude many redundant, but relevant, variables. We present in Section 1.2 a short review of the literature employed for variable selection and in Section 1.3 we introduce a more formal approach to the problem.

Consider a set of m observations $\{x_k, y_k\}$,$(k = 1, ..., m)$ consisting of n input variables $x_{k,i}$, $(i = 1, ..., n)$ and one output variable y_k. Variable ranking makes use of a scoring function $S(i)$ computed from the values $x_{k,i}$ and y_k, $k = 1, ..., m$. By convention, we assume that a high score is indicative of a valuable variable and that we sort variables in decreasing order of $S(i)$. To use variable ranking to build predictors, nested subsets incorporating progressively more and more variables of decreasing relevance are defined. Following the classification of Kohavi and John (1997), variable ranking is a filter method: it is a pre-processing step, independent of the choice of the predictor. Still, under certain independence or orthogonal assumptions, it may be optimal with respect to a given predictor. For instance, using Fisher's criterion to rank variables in a classification problem where the covariance matrix is diagonal is optimal for Fisher's linear discriminant classifier (Duda et al., 2001). Even when variable ranking is not optimal, it may be preferable to other variable subset selection methods because of its computational and statistical scalability: computationally, it is efficient since it requires only the computation of n scores and sorting the scores; statistically, it is robust against overfitting because it introduces bias but it may have considerably less variance (Hastie et al., 2001).

If the input vector x can be interpreted as the realization of a random vector drawn from an underlying unknown distribution, we denote by X_i the random variable corresponding to the $i-th$ component of x. Similarly, Y will be the random variable of which the outcome y is a realization. We further denote by x_i the m dimensional vector containing all the realizations of the $i-th$ variable for the training observations, and by y the m dimensional vector containing all the target values.

In the case where there is a large number of variables that separate the data perfectly, ranking criteria based on classification success rate cannot distinguish between the top ranking variables.

Several approaches to the variable selection problem using information theoretic criteria have been proposed (as reviewed e.g. by Bekkerman et al., 2003, Dhillon et al., 2003, Forman, 2003, Torkkola, 2003). Many rely on empirical estimates of the

mutual information between each variable and the target:

$$I(i) = \int_{x_i} \int_y p(x_i, y) log \frac{p(x_i, y)}{p(x_i)p(y)} dxdy, \tag{1.1}$$

where $p(x_i)$ and $p(y)$ are the probability densities of x_i and y, and $p(x_i, y)$ is the joint density. The criterion $I(i)$ is a measure of dependency between the density of the variable x_i and the density of the target y. The difficulty is that the densities $p(x_i)$, $p(y)$ and $p(x_i, y)$ are all unknown and are hard to estimate from the data. The case of discrete or nominal variables is probably the easiest because the integral becomes a sum:

$$I(i) = \sum_{x_i} \sum_y P(X = x_i, Y = y) log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)}. \tag{1.2}$$

The probabilities are then estimated from frequency counts. For example, in a three-class problem, if a variable takes 4 values, $P(Y = y)$ represents the class frequency (3 frequency counts), $P(X = x_i)$ represents the distribution of the input variable (4 frequency counts), and $P(X = x_i; Y = y)$ is the frequency of the joint observations (12 frequency counts). The estimation obviously becomes harder with larger numbers of classes and variable values. The case of continuous variables (and possibly continuous targets) is the hardest. One can consider discretizing the variables or approximating their densities with a non-parametric method such as Parzen windows (see, e.g. Torkkola, 2003).

Variable subset selection approaches essentially divide into wrappers, filters, and embedded methods. Wrappers utilize the learning model of interest as a black box to score subsets of variables according to their predictive power. Filters select subsets of variables as a pre-processing step, independently of the chosen predictor. Embedded methods perform variable selection in the process of training and are usually specific to given learning methods.

The wrapper methodology, recently popularized by Kohavi and John (1997), offers a simple and powerful way to address the problem of variable selection, regardless of the chosen learning model. In its most general formulation, the wrapper methodology consists in using the prediction performance of a given learning model to assess the relative usefulness of subsets of variables. The problem is known to be NP-hard (Amaldi and Kann, 1998) and the search becomes quickly computationally intractable. A wide range of search strategies can be used, including best-first,

branch-and-bound, simulated annealing, genetic algorithms (see Kohavi and John, 1997, for a review). Performance assessments are usually done using a validation set or by cross-validation. Possible predictors include decision trees, naive Bayes, least-square linear predictors, and support vector machines. Wrappers are often criticized because they seem to be a "brute force" method requiring massive amounts of computation, but it is not necessarily so. Efficient search strategies may be devised; using such strategies does not necessarily mean sacrificing prediction performance. Greedy search strategies seem to be particularly computationally advantageous and robust against overfitting. They come in two flavours: forward selection and backward elimination. In forward selection, variables are progressively incorporated into larger and larger subsets, whereas in backward elimination one starts with the set of all variables and progressively eliminates the least promising ones. Both methods yield nested subsets of variables. Wrapper methods are remarkably universal and simple. However, embedded methods that incorporate variable selection as part of the training process may be more efficient in several respects: they make better use of the available data by not needing to split the training data into a training and validation set; they reach a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated. Embedded methods are not new: decision trees such as CART, for instance, have a built-in mechanism to perform variable selection (Breiman et al., 1984).

Some embedded methods guide their search by estimating changes in the objective function value incurred by making moves in variable subset space. Combined with greedy search strategies (backward elimination or forward selection) they yield nested subsets of variables.

A lot of progress has been made to formalize the objective function of variable selection and algorithms to optimize it. Generally, the objective function consists of two terms that compete with each other: (1) the goodness-of-fit (to be maximized), and (2) the number of variables (to be minimized). This correspondence is formally established in the paper of Weston et al. (2003) for the particular case of classification with linear predictors. Weston et al. note that, although their algorithm only approximately minimizes the l_0-norm in practice it may generalize better than an algorithm that really did minimize the "0-norm", because the latter would not provide sufficient regularization (a lot of variance remains because the optimization problem

has multiple solutions). The need for additional regularization is also stressed in the paper of Perkins et al. (2003).

To our knowledge, no algorithm has been proposed to directly minimize the number of variables for non-linear predictors. Instead, several authors have substituted for the problem of variable selection that of variable scaling (Weston et al., 2000, Grandvalet and Canu, 2002). The variable scaling factors are "hyper-parameters" adjusted by model selection. The scaling factors obtained are used to assess variable relevance. A variant of the method consists of adjusting the scaling factors by gradient descent on a bound of the leave-one-out error (Weston et al., 2000). This method is used as baseline method in the paper of Weston et al. (2003).

In some applications, reducing the dimensionality of the data by selecting a subset of the original variables may be advantageous for reasons including the expense of making, storing and processing measurements. If these considerations are not of concern, other means of space dimensionality reduction should also be considered.

Clustering, the most important unsupervised learning method, has long been used for feature construction. The idea is to replace a group of "similar" variables by a cluster centroid, which becomes a feature. The most popular algorithms include K-means and hierarchical clustering. For a review, see, e.g., the textbook of Duda et al. (2001) and Giudici (2003). Clustering is usually associated with the idea of unsupervised learning. It can be useful to introduce some supervision in the clustering procedure to obtain more discriminant features. This is the idea of distributional clustering (Pereira et al., 1993). Distributional clustering is rooted in the information bottleneck (IB) theory of Tishby et al. (1999).

Another widely used method of feature construction is singular value decomposition (SVD). The goal of SVD is to form a set of features that are linear combinations of the original variables, which provide the best possible reconstruction of the original data in the least square sense (Duda et al., 2001). It is an unsupervised method of feature construction. There are results on information theoretic unsupervised feature construction method: sufficient dimensionality reduction (SDR). The most informative features are extracted by solving an optimization problem that monitors the trade-off between data reconstruction and data compression, similar to the information bottleneck of Tishby et al. (1999); the features are found as Lagrange multipliers of the objective optimized.

1.3 A general framework for variable selection

The objective of variable selection can be started as how to select good variables according to a maximal statistical dependency criterion based on mutual information. Because of the difficulty in directly implementing the maximal dependency condition, we first derive an equivalent form, called minimal-redundancy-maximal-relevance criterion (mRMR), for first-order incremental feature selection. Then, we present a two-stage feature selection by combining mRMR and other more sophisticated feature selectors. Given the input data D tabled as N samples and M features, $X = (x_i, i = 1, ..., M)$, and the target classification variable c, the feature selection problem is to find from an M-dimensional observation space,R^M a subspace of m features, R^m, that "optimally" characterizes c. Given a condition defining the "optimal characterization" an algorithm is needed to find the best subspace. The optimal characterization condition often means the minimal classification error. In an unsupervised situation where the classifiers are not specified, minimal error usually requires the maximal statistical dependency of the target class c on the data distribution in the subspace R^m (and vice versa). This scheme is *Maximal Dependency* . One of the most popular approaches to realize Max-Dependency is *Maximal Relevance* feature selection: selecting the features with the highest relevance to the target class c. Relevance is usually characterized in terms of correlation or mutual information, of which the latter is one of the widely used measures to define dependency of variables.

Here, we focus on the discussion of mutual-information-based feature selection. Given two random variables X and Y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(x)$, and $p(x, y)$:

$$I(i) = \int_x \int_y p(x, y) log \frac{p(x, y)}{p(x)p(y)} dxdy. \tag{1.3}$$

In Max-Relevance, the selected features x_i are required, individually, to have the largest mutual information $I(x_i; c)$ with the target class c, reflecting the largest dependency on the target class. In terms of sequential search, the m best individual features, i.e., the top m features in the descent ordering of $I(x_i, c)$, are often selected as the m features. In variable selection, it has been recognized that the combinations of individually good features do not necessarily lead to good classification performance. In other words, "the m best features are not the best m features", see e.g.

Cover and Thomas (1991).

Here we present an empirical minimal-redundancy-maximal-relevance (mRMR) framework to minimize redundancy, and use a series of intuitive measures of relevance and redundancy to select promising features for both continuous and discrete data sets. First, although both Max- Relevance and Min-Redundancy have been intuitively used for feature selection, no theoretical analysis is given on why they can benefit selecting optimal features for classification. Thus, the first goal of this dissertation is to present a theoretical analysis showing that mRMR is equivalent to Max-Dependency for first-order feature selection, but is more efficient.

1.3.1 Relationships of Max-Dependency, Max-Relevance and Min-redundancy

In terms of mutual information, the purpose of feature selection is to find a feature set S with m features (x_i) which jointly have the largest dependency on the target class c. This scheme, called Max-Dependency, has the following form:

$$maxD(S,c), D = I\left\{(x_i, i = 1, ..., m); c\right\} \tag{1.4}$$

Obviously, when m equals 1, the solution is the feature that maximizes $I(x_j; c), (1 \leq j \leq M)$. When $m > 1$ a simple incremental search scheme is to add one feature at one time: given the set with $m-1$ features, S_{m-1}, the m-th feature can be determined as the one that contributes to the largest increase of $I(S; c)$, which takes the form of:

$$\begin{aligned} I(S_m; c) &= \int\int p(S_m, c) log \frac{p(S_m, c)}{p(S_m)p(c)} dS_m dc \\ &= \int\int p(S_{m-1}, x_m, c) log \frac{p(S_{m-1}, x_m, c)}{p(S_{m-1}, x_m)p(c)} dS_{m-1} dx_m dc \\ &= \int\int p(x_1, ..., x_m, c) log \frac{p(x_1, ..., x_m, c)}{p(x_1, ..., x_m)p(c)} dx_1 ... dx_m dc \end{aligned} \tag{1.5}$$

Despite the theoretical value of Max-Dependency, it is often hard to get an accurate estimation for the multivariate density $p(x_1, ..., x_m)$ and $p(x_1, ..., x_m, c)$ because of two difficulties in the high-dimensional space: 1) the number of samples is often insufficient and 2) the multivariate density estimation often involves computing the inverse of the high-dimensional covariance matrix, which is usually an ill posed problem.

Another drawback of Max-Dependency is the slow computational speed. These problems are most pronounced for continuous feature variables. Even for discrete (categorical) features, the practical problems in implementing Max-Dependency cannot be completely avoided. For example, suppose each feature has three categorical states and N samples. K features could have a maximum $min(3^k, N)$ joint states. When the number of joint states increases very quickly and gets comparable to the number of samples, N, the joint probability of these features, as well as the mutual information, cannot be estimated correctly. Hence, although Max-Dependency feature selection might be useful to select a very small number of features when N is large, it is not appropriate for applications where the aim is to achieve high classification accuracy with a reasonably compact set of features.

As Max-Dependency criterion is hard to implement, an alternative is to select features based on maximal relevance criterion (Max-Relevance). Max-Relevance is to search features which approximates $D(S, c)$ with the mean value of all mutual information values between individual features x_i and class c:

$$maxD(S, c), D = \frac{1}{dim(S)} \sum_{x_i \in S} I(x_i; c). \tag{1.6}$$

It is likely that features selected according to Max- Relevance could have rich redundancy, i.e., the dependency among these features could be large. When two features highly depend on each other, the respective class-discriminative power would not change much if one of them were removed. Therefore, the following minimal redundancy (Min-Redundancy) condition can be added to select mutually exclusive features:

$$minR(s), R = \frac{1}{dim(S)^2} \sum_{x_i, x_j \in S} I(x_i, x_j). \tag{1.7}$$

The criterion combining the above two constraints is called "minimal-redundancy-maximal-relevance",(mRMR). We define the operator $\phi(D, R)$ to combine to combine D and R and consider the following simplest form to optimize D and R simultaneously:

$$max\phi(D, R), \phi = D - R. \tag{1.8}$$

In practice, incremental search methods can be used to find the near optimal features defined by $\phi()$. Suppose we already have S_{m-1}, the feature set with $m-1$ features. The task is to select the m-th feature from the set $X - S_{m-1}$. This is done by

selecting the feature that maximizes $\phi()$. The respective incremental algorithm optimizes the following condition:

$$max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right]. \tag{1.9}$$

The computational complexity of this incremental search method is $O(dim(S) \times M)$. We propose and we prove now the following results:

Lemma: *The combination of Max- Relevance and Min-Redundancy criteria, (the mRMR criterion), is equivalent to the Max-Dependency criterion if one feature is selected (added) at one time.*

Here we give an idea for the proof.

We assume that S_{m-1} i.e., the set of $m-1$ features, has already been obtained. The task is to select the optimal $m-th$ feature x_m from set $X - S_{m-1}$.

The dependency D is represented by mutual information, i.e., $D = I(S_m; c)$, where $S_m = S_{m-1}, x_m$ can be treated as a multivariate variable. Thus, by the definition of mutual information, we have:

$$\begin{aligned} I(S_m; c) &= H(c) + H(S_m) - H(S_m, c) \\ &= H(c) + H(S_{m-1}, x_m) - H(S_{m-1}, x_m, c), \end{aligned} \tag{1.10}$$

where $H(.)$ is the entropy of the respective multivariate (or univariate) variables. Now, we define the following quantity $J(S_m) = J(x_1, ..., x_m)$ for scalar variables $x_1, ..., x_m$,

$$J(x_1, ..., x_m) = \int ... \int p(x_1, ..., x_m) log \frac{p(x_1, x_2, ..., x_m)}{p(x_1)...p(x_m)} dx_1 ... dx_m. \tag{1.11}$$

Similarly, we define $J(S_m, c) = J(x_1, ..., x_m, c)$ as

$$J(x_1, ..., x_m, c) = \int ... \int p(x_1, ..., x_m, c) log \frac{p(x_1, x_2, ..., x_m, c)}{p(x_1)...p(x_m)p(c)} dx_1 ... dx_m dc. \tag{1.12}$$

We can easily derive from the previous equations:

$$H(S_{m-1}, x_m) = H(S_m) = \sum_{i=1}^{m} H(x_i) - J(S_m), \tag{1.13}$$

and

$$H(S_{m-1}, x_m, c) = H(S_m, c) = H(c) + \sum_{i=1}^{m} H(x_i) - J(S_m, c). \tag{1.14}$$

By substituting them to the corresponding terms in $I(S_m, c)$, we have:

$$\begin{aligned} I(S_m; c) &= J(S_m, c) - J(S_m) \\ &= J(S_{m-1}, x_m, c) - J(S_{m-1}, x_m). \end{aligned} \tag{1.15}$$

Obviously, Max-Dependency is equivalent to simultaneously maximizing the first term and minimizing the second term. We can use Jensen's Inequality to show the second term $J(S_{m-1}, x_m)$ is lower-bounded by 0. A related and slightly simpler proof is to consider the inequality $log(z) \leq z - 1$ with the equality if and only if $z = 1$. We see that:

$$\begin{aligned} &-J(x_1, ..., x_m) = \\ &= \int ... \int p(x_1, ..., x_m) log \frac{p(x_1)...p(x_m)}{p(x_1, ..., x_m)} dx_1...dx_m \\ &\leq \int ... \int p(x_1, ..., x_m) \left[\frac{p(x_1)...p(x_m)}{p(x_1, ..., x_m)} - 1 \right] dx_1...dx_m \\ &= \int ... \int p(x_1)...p(x_m) dx_1...dx_m - \int ... \int p(x_1, ..., x_m) dx_1...dx_m \\ &= 1 - 1 = 0 \end{aligned} \tag{1.16}$$

It is easy to verify that the minimum is attained when $p(x_1, ..., x_m) = \prod_{i=1}^{m} p(x_i)$,i.e., all the variables are independent of each other.

As all the $m - 1$ features have been selected, this pair-wise independence condition means that the mutual information between x_m and any selected feature $x_i, i = 1, ..., m - 1$ is minimized. This is the Min-Redundancy criterion. We can also derive the upper bound of the first term in $J(S_{m-1}, c, x_m)$.

For simplicity, let us first show the upper bound of the general form $J(y_1, ...y_n)$, assuming there are n variables $y_1, ..., y_n$. This can be seen as follows:

$$\begin{aligned} &J(y_1, ..., y_n) = \\ &= \int ... \int p(y_1, ..., y_n) log \frac{p(y_1...y_n)}{p(y_1), ..., p(y_n)} dy_1...dy_n \\ &= \int ... \int p(y_1, ..., y_n) log \frac{p(y_1|y_2...y_n)...p(y_{n-1}|y_n p(y_n)}{p(y_1)...p(y_n)} dy_1...dy_n \\ &= \sum_{i=1}^{n-1} H(y_i) - H(y_1|y_2, ..., y_n) - ... - H(y_{n-1}|y_n) \\ &\leq \sum_{i=1}^{n-1} H(y_i). \end{aligned} \tag{1.17}$$

This equation can be easily extended as:

$$J(y_1, ..., y_n) \leq min \tag{1.18}$$

$$\left\{ \sum_{i=2}^{n} H(y_i), \sum_{i=1, i\neq 2}^{n} H(y_i), ..., \sum_{i=1, i\neq n-1}^{n} H(y_i), \sum_{i=1}^{n-1} H(y_i) \right\}.$$

It is easy to verify the maximum of $J(y_1, ..., y_n)$ or, similarly $J(S_{m-1}, c, x_m)$, is attained when all variables are maximally dependent. When S_{m-1} has been fixed, this indicates that x_m and c should have the maximal dependency. This is the Max-Relevance criterion.

Therefore, a combination of Max- Relevance and Min-Redundancy is equivalent to Max-Dependency for first-order selection. We have the following observations:

- Minimizing $J(S_m)$ only is equivalent to searching mutually exclusive (independent) features. This is insufficient for selecting highly discriminative features.
- Maximizing $J(S_m, c)$ only leads to Max-Relevance. Clearly, the difference between mRMR and Max- Relevance is rooted in the different definitions of dependency (in terms of mutual information); does not consider the joint effect of features on the target class. On the contrary, Max-Dependency considers the dependency between the data distribution in subspace R^m and the target class c. This difference is critical in many circumstances.

1.3.2 Computational Issues

We consider mutual-information-based feature selection for both discrete and continuous data. For discrete (categorical) feature variables, the integral operation reduces to summation. In this case, computing mutual information is straightforward, because both joint and marginal probability tables can be estimated from the samples of categorical variables in the data.

However, when at least one of variables x and y is continuous, their mutual information $I(x, y)$ is hard to compute, because it is often difficult to compute the integral in the continuous space based on a limited number of samples.

One solution is to incorporate data discretization as a pre-processing step. For some applications where it is unclear how to properly discretize the continuous data, an alternative solution is to use density estimation method (e.g., Parzen windows) to approximate $I(x, y)$,as suggested by earlier work in medical image registration and feature selection.

Given N samples of a variable x, the approximate density function $\hat{p}(x)$ has the

following form:

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{n}\delta(x - x^{(i)}, h), \tag{1.19}$$

where $\delta()$ is the Parzen window function as explained below, $x^{(i)}$ is the $i-th$ sample, and h is the window width. Parzen has proven that, with the properly chosen $\delta()$ and h, the estimation $\hat{p}(x)$ can converge to the true density $p(x)$ when N goes to infinity. Usually, $\delta()$ is chosen as the Gaussian window:

$$\delta(z, h) = exp\left(-\frac{z^T\Sigma^{-1}}{2h^2}\right) / \left\{(2\pi)^{d/2}h^d|\Sigma|^{1/2}\right\}, \tag{1.20}$$

where $z = x - x^{(i)}$, d is the dimension of the sample x and Σ is the covariance of z. When $d = 1$, the previous equation returns the estimated marginal density; when $d = 2$,we can use previous equation to estimate the density of bivariate variable. For the sake of robust estimation, for $d \geq 2$, Σ is often approximated by its diagonal components.

It is possible also to extend our proposal in a Bayesian framework. The idea is to employ the Bayes rule and assuming that feature variables are independent of each other given the target class. Given a sample, $s = (x_1, ..., x_m)$ for m features, the posterior probability that s belongs to class c_k is:

$$p(c_k|s) \propto \prod_{i=1}^{m} p(x_i|c_k), \tag{1.21}$$

where $p(x_i|c_k)$ is the conditional probability table (or densities) learned from examples in the training process. The Parzen-window density-approximation can be used to estimate $p(x_i|c_k)$ for continuous features and, therefore, the posterior class probabilities are obtained.

Chapter 2

Lifetime value models

In this work we consider variable selection methods for Lifetime Value Models. Customer Lifetime Value (LTV) measures the profit generating potential, or value, of a customer, and is increasingly being considered a touchstone for administering Customer Relationship Management (CRM) process. This in order to provide attractive benefits and retain high value customers, while maximizing profits from a business standpoint. Robust and accurate techniques for modelling LTV are essential in order to facilitate CRM via LTV. A customer LTV model needs to be explained and understood to a large degree before it can be adopted to facilitate CRM. LTV is usually considered to be composed of two independent components: tenure and value. Though modelling the value (or equivalently, profit) component of LTV, (which takes into account revenue, fixed and variable costs), is a challenge in itself, our experience has revealed that finance departments, to a large degree, well manage this aspect. Therefore, in this paper, our focus will mainly be on modelling tenure rather than value.

2.1 Customer Lifetime Value: introduction

In the medical world, doctors often want to understand which treatments help patients survive longer and which have no effect at all. In the business world, the equivalent concern is on customers lifetime. This is particularly true of businesses that have a well defined beginning and end to the customer relationship subscription based relationships. These relationships are found in a wide range of industries,

such as insurance, communication, cable televisions, newspaper magazine subscription, banking, and electricity providers in competitive markets.

The basis of survival analysis is the hazard probability, the chance that someone who has survived for a certain length of time (called the customer tenure) is going to stop, cancel, or expire before the next unit of time. This definition assumes that time is discrete, and such discrete time intervals, whether days, weeks, or months, often fits business needs. By contrast, traditional survival analysis in statistics usually assumes that time is continuous.

Given the right data, calculating the hazard probability for a given tenure t is simple. The probability is the number who succumbed to the risk divided by the population at risk at that tenure. That is, the numerator is the number of customers who stopped with exactly tenure t and the denominator is everyone who had tenures greater than or equal to t. Customers with shorter tenures are not part of the risk group.

Figure 2.1 charts hazard probabilities for customers in a typical subscription business. The horizontal axis is the tenure of customers measured in days; the vertical axis is the probability that customers stop at a particular tenure point. The hazard

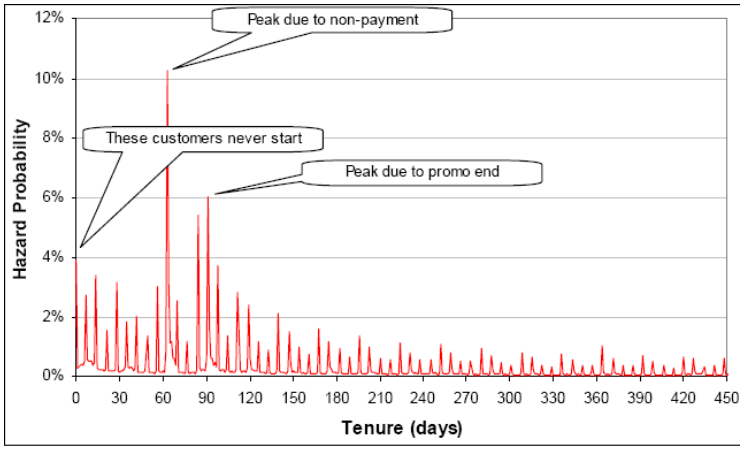

Figure 2.1: Customer hazard probabilities

chart is an X ray into the customer lifecycle, because it highlights different important events. The very first hazard probability at time zero is about 0.04; this is due to customers not starting and is often caused by poor customer information being gathered at the point of sale or perhaps by buyer remorse. Around 60 days, there is

a very strong peak in the hazard probability. This corresponds to those customers who start but never pay. The company moves customers through various dunning levels to inspire payment. However, at some point, the company must force churn because of non-payment. Changes in this policy, such as a reduction in the period of time for cutting off non-paying customers, would be apparent in the hazard probabilities. Around 90 days, there is another significant spike in the hazards in Figure 2.1. This spike actually has nothing to do with non-payment. It is due to the end of the initial promotion. Customers who sign up for this service because the initial offer is cheap often stop when they have to start paying full price. Happily, the customers who stop at this point have at least been paying their bills. After these two initial peaks, the hazard probability gradually declines but with a jagged characteristic. The jaggedness is actually due to the one-month billing cycle that most customers are on. Customers are more likely to stop at the end of a billing cycle. One reason is that when customers call in to stop, the stop date is set to the end of the billing cycle unless the customer requests a specific date. The gradual decline in hazards is also interesting. In fact, it says something quite important about customer loyalty: The longer customers stay with the company, the less likely they are to leave. The long term decline in hazards is as good a measure of loyalty that we know.

If hazard curves provide an X ray into the customer lifecycle, survival curves provide a more holistic picture. The survival at time t is simply the likelihood that a customer will survive to that point in time. This is calculated directly from the hazards, by taking the cumulative probability that someone does not stop before time t, that is, by multiplying one minus the hazards together for all values less than t.

Survival does more than show the difference between groups of customers. It makes it possible to quantify the difference between groups. The chart in Figure 2.2 illustrates one common measure, the customer half-life (or median customer lifetime). This is the tenure where exactly half the original customers would still be expected to be active. The calculation is quite easy. The vertical axis has the survival values. Figure 2.2 shows the median tenure by payment type for the groups shown earlier. For credit card payers, the median is over 240 days; for others, barely a quarter that. The customer half life provides a good way to compare different groups of customers. One drawback to the customer half life is that the survival curve may not cross 0.50. This means that the customer half life is not known, because the time window is not

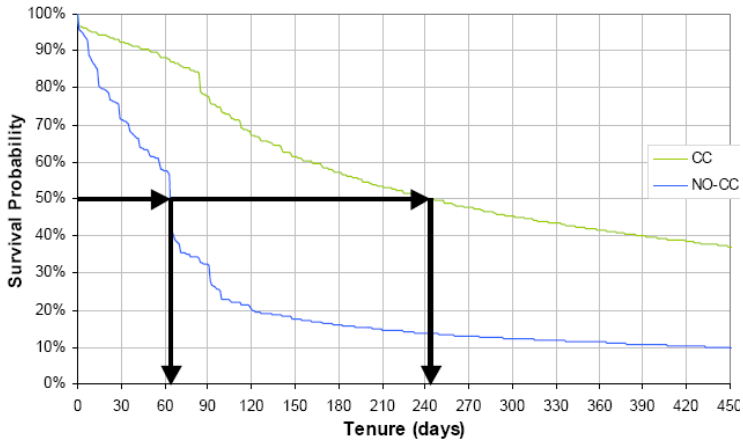

Figure 2.2: Customer hazard probabilities

large enough. Extrapolating survival beyond the time window is dangerous because what happens to customers is not known. Customers may stay around for another hundred years or they might all stop the next day.

The customer half life is a good comparison for different groups of customers. However, this value only tells us about one customer, the customer whose tenure is exactly in the middle. A more useful number is the average customer lifetime, which can be dropped directly into customer value calculations. If a subscription is worth 500 per year in revenue and the average customer lifetime is 2.5 years, then the customer is worth $1,250$ (assuming no discounting of future revenue).

Calculating the average tenure is conceptually quite easy. It turns out that the average tenure for a given period of time is the area under the survival curve. For instance, the average tenure in the first year after acquisition for customers who stop half way through the year is half a year. On the other hand, customers who survive longer than one year only get one year, because the calculation is only looking at the first year tenure. The average for all customers is the area under the survival curve up to 365 days.

Another critical idea in survival analysis is that of competing risks. When studying the survival rates for cancer victims, what happens when someone enrolled in the study dies in a car accident? Or moves to a foreign country? In medical terminology, these patients are "lost to follow up". The same thing can happen with customers. A clear example of competing risks is the distinction between voluntary

and involuntary churn. Some customers are forced to leave (typically due to non payment) whereas others leave voluntarily. When looking at churn, sometimes models are built leaving out one or the other group of customers. However, this results in a biased model one of the issues when developing payment risk models separate from voluntary churn models. With competing risks, the approach to the problem is a bit different. Customers who voluntarily stopped at a particular tenure, say one year, did not stop either voluntarily or involuntarily before then. This is useful information for understanding both types of churn.

Calculating competing risks follows the same pattern described earlier with one difference. For each tenure, there is a separate probability for each risk; once a customer has succumbed to one risk (say voluntary churn), the customer is no longer included in the population at risk for any of the risk groups. Technically, the customer is censored for other risks. Figure 2.3 shows competing risks for voluntary and involuntary churn for the credit card paying and non credit card customers shown earlier. The top line shows clearly that credit card paying customers are at minimal risk for involuntary churn. Although the canonical example is voluntary versus

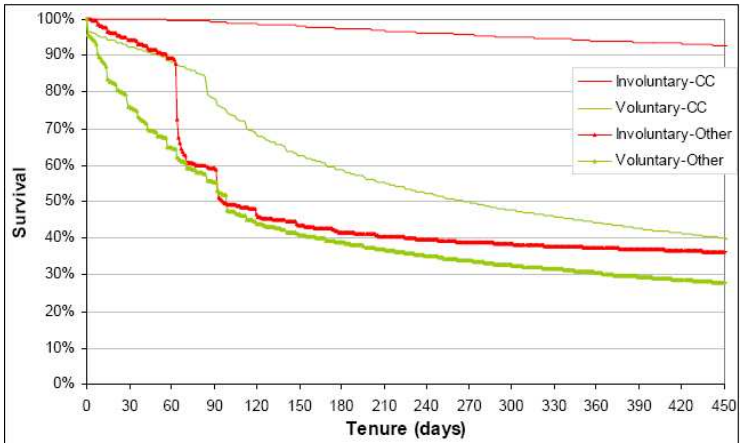

Figure 2.3: Types of churn

involuntary churn, competing risks is useful in other situations. For instance, some customers may "churn" because they migrate to a higher value product. A wireless customer upgrading to more advanced technology may count as "churn" on the old technology. A cable subscriber who switches to digital cable may count as "churn" on her previous account. This suggests including migration as a competing risk for

understanding this customers.
In the next Section we introduce the theoretical problem about customer lifetime estimation.

2.2 Customer Lifetime Value: a mathematical approach

Historical data, extracted from operational customer databases, can be used to build predictive models for various temporal outcomes: cancellation of products or services (churn), downgrading, acquiring add-on products or upgrading, product return, and loan prepayment.
The occurrence of the target event on the $i-th$ customer is controlled by the probability distribution of the time until the event, T_i. Customer events might be recorded at discrete increments such as months or on a continuous time scale. At the time the data was extracted for analysis, all customers usually have not experienced the event. In this case, the event time is considered (right) censored. Survival analysis is a set of statistical methods designed for censored duration data. Censored event history data can be represented by an observed event time, $Y_i = min(T_i, a - B_i)$, and an event indicator, $\delta_i = I\{T_i \leq a - B_i\}$. The date of origin, B_i can vary among customers. Typically, B_i represents the date that an account was opened. In this censoring scheme (generalized type I censoring), there is a fixed date, a, when the extracted data was current, see Figure 2.4. Another possible cause of censoring is the occurrence of an independent and mutually exclusive competing event. For example, if the event of interest is cancellation of a service, then a customer that moves out of the service area might be considered censored at the date they moved, a_i. The data used for mining customer histories consists of retrospective samples extracted from large operational databases. With discrete event times the truncation date can equal the censoring date. The available data might be all accounts active at the beginning of the month, some of which experienced the event during the month. In Figure 2.4, a line segment represents each subject. The vertical axis is the event time. The horizontal axis is the calendar date. The beginning of each line segment corresponds to the origin $(B_i, 0)$. The end of the line segment corresponds to the event time (B_i+T_i, T_i). The eight subjects with lines extending beyond the censoring date, $a = 1000$, are censored at $Y_i = 1000 - B_i$. If the sample were truncated at

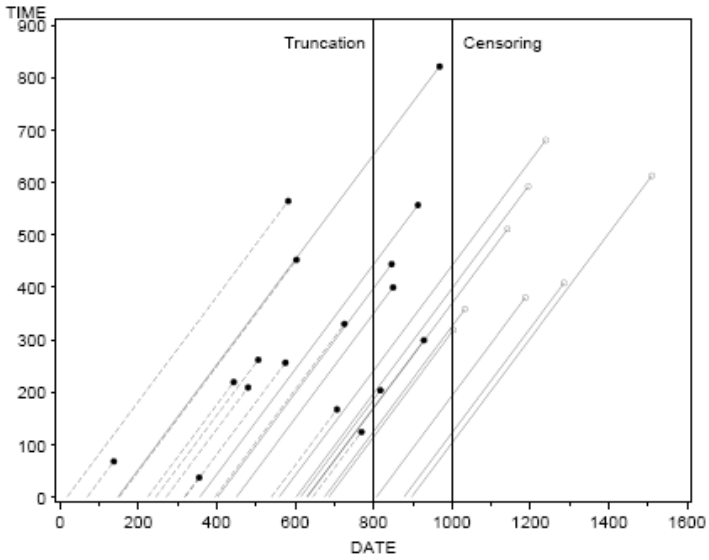

Figure 2.4: Censored event history

the date $c = 800$, the 11 dashed lines would be absent. The event time distribution is usually characterized by the survival function, or the hazard rate. For discrete event times, the hazard rate is the conditional probability of the event given that it has yet to occur

$$h(t_j) = Pr(T = t_j | T \geq t_j) = \frac{1 - S(t_{j+1})}{S(t_j)}. \tag{2.1}$$

The hazard can be interpreted as an age-specific rate (events/unit time). The survival function decreases monotonically from one to zero. In contrast, the hazard rate can be any nonnegative function. The shape of the hazard rate often gives insight into the underlying system driving the occurrence of an event Customer databases contain concomitant information that may affect the event time distribution such as demographics, account balances and payments, and the occurrence of other events such as the acquisition of new products or services. The vector of covariates for the $i - th$ customer is often time-dependent. Time-dependent covariates can represent single irreversible events that occur at some point in the customer lifetime such as paying off an instalment loan. Time-dependent covariates can be step functions representing the occurrences of repeatable events such as problems reported to customer service or payment delinquencies. Time-dependent covariates can be more continuously varying quantities such as the balance in an investment account.

The ultimate purpose of modelling customer relationship data is usually prediction. Predictive models are used to map attributes of each customer to a score, which measures the propensity of some actionable event. The choice of an appropriate predictive score depends on how the model is to be deployed. In the most general scenario, customers would be scored at the current point in their lifetime for the propensity of the outcome. Consequently, predictive scoring should consider the distribution of the residual event time $R = (T - t|T \geq t)$. The hypothetical random variable R is the time remaining until the event, conditional on the information available at the current time t. The hazard rate at t, $h(t)$, equals the probability density (mass) function of R and can be interpreted as the probability of the event in the next instant. The hazard rate is a relevant score in many applications. Other potentially useful scores can be derived from the distribution of the residual event time. The expected value of R (mean residual life) is the area under survival function of R:

$$\mu(t) = E(T - t|T \geq t) = \frac{1}{S(t)} \int_{\infty}^{t} S(x)dx. \tag{2.2}$$

The median of R (median residual life) is the half-life of the remaining time

$$m(t) = med(T - t|T \geq t) = S^{-1}\left(\frac{1}{2}S(t)\right) - t \tag{2.3}$$

The mean and median are defined similarly for discrete event times. Smaller quantiles of the residual event time (e.g., quarter-life) can be useful with heavily censored data because the mean and median can fall far outside the range of observed events. If the model is used to score new customers at time zero, then $\mu(0)$ and $m(0)$ revert to the mean and median of T. In many cases, the scores are used to forecast a future time $t + r$ The probability density function of R evaluated at r is more relevant than the hazard rate at $t + r$ because T is only known to be greater than t:

$$f_R(r) = \frac{f(t + r)}{S(t)}. \tag{2.4}$$

In forecasting applications, the time-dependent covariates would either need to be forecasted or lagged by r units in the model. When the entire future interval $[t, t+r]$ is of interest, the survival function of R evaluated at r is pertinent

$$S_R(r) = Pr(T \geq t + r|T \geq t) = \frac{S(t + r)}{S(r)}. \tag{2.5}$$

This quantity is monotonically related to the cumulative hazard (total risk) on the interval $[t, t + r]$ The area under the survival function of R on the interval $[t, t + r]$

involves all values in the interval, not just the endpoints. This area is equal to the restricted mean residual event time:

$$\mu(t,r) = E(\min(R,r)) = \frac{1}{S(t)} \int_t^{t+r} S(x)dx. \tag{2.6}$$

When $t = 0$, the restricted residual event time is the restricted mean life.
An ideal predictive scoring model would give a sufficiently flexible estimate of the hazard rate as a function of the (possibly time-dependent) covariates. The discrete-time logistic hazard model and the piecewise exponential hazard model passably satisfy this requirement. The hazard rate uniquely characterizes the event time distribution. The survival function can be determined from the hazard rate using the identities:

$$S(t) = exp\left(-\int_0^t h(x)dx\right) \tag{2.7}$$

and

$$S(t) = \prod_{t_j<t} (1 - h(t_j)) \tag{2.8}$$

for continuous and discrete times, respectively.

2.2.1 Predicting Customer Value

In this Section we want to describe how to calculate Customer value. Customer lifetime value (CLV) is used for allocating discretionary marketing investments and for corporate valuation (Bauer et al. 2003). Pfeifer et al. (2004) distinguish the finance usage of value from the accounting usage of profit.
Our proposal: a customer lifetime value, in our connotation, is the present value of customers future cash inflows minus cash outflows. In contrast, customer profitability is their accrual based revenue minus costs over a fixed, usually past, time. The remaining value of a customer from the current time t forward depends on their residual life $R = 0, 1, 2, 3, ...$
The residual life is the time remaining until churn, the end of the customer relationship. It is a discrete random variable having non-negative integer values. Zero means that the customer churned at the current time t. If the current time is zero the customer just started, then their residual life is their entire life. The remaining CLV from the current time t forward is the sum of the discounted cash flows over

the residual life:

$$val(R_t) = \sum_{r=0}^{R_t} \delta_r (1+d)^{-r}. \tag{2.9}$$

The cash flows δ_r linked to the customer relationship can vary in time. The churn time is defined to be the last time with non-zero cash flows attributed to the customer. Thus, all customers active at t are worth at least δ_0, even those who churned at $R_t = 0$.

The present value of future cash flows is discounted for not having the opportunity to invest the money at time t. The discount factor $(1+d)^{-r}$ is the amount that would be paid now for one monetary unit received at one time later. The discount rate per time unit d is compounded once every time unit.

CLV is a random variable. Business decisions are based on the centre of its probability distribution. The general formula for the expected CLV is an infinite sum of an ever-expanding product:

$$\begin{aligned} E(val(R_t)) &= \sum_{r=0}^{\infty} val(r) Pr(R_t = r) \\ &= \sum_{r=0}^{\infty} \sum_{j=0}^{r} \delta_j (1+d)^{-j} Pr(R_t = r) \\ &= \delta_0 + \sum_{r=1}^{\infty} \delta_r (1+d)^{-r} \prod_{j=t}^{t+r-1} (1 - h(j|x(j))). \end{aligned} \tag{2.10}$$

For discrete-time data, the hazard function $h(t)$ is the conditional probability of churn at time t given that the customer has not churned yet.

$$h(t|x(j)) = Pr(T = t | T \geq t, x(j)) = Pr(R_t = 0 | x(j)). \tag{2.11}$$

The random variable T is the duration until churn (customer tenure). The hazards depend on a vector of customer covariates $x(t)$.

Note that the mean CLV does not equal the value function evaluated at the mean residual life:

$$E(val(R_t)) \neq \sum_{r=0}^{E(R_t)} \delta_r (1+d)^{-r}, \tag{2.12}$$

(unless the cash flows are constant and the discount rate is zero). CLV is a nonlinear function of the residual life, so simply plugging in the mean of R_t R gives only a rough first-order approximation. The formula for the mean CLV depends on three unknowns:

- The value function needs to be specified;

- The churn hazard needs to be modelled as a function of tenure and other customer covariates;
- The summation goes to infinity, but the data does not. Either we can make assumptions about the shape of the hazard outside the range of the data or we need to reformulate the problem.

2.2.2 Specifying the value function

If we assume that the δ_r are time-constant, then they can be estimated using the current revenue and expenses associated with the customer products and services. This accounting could be done for each individual or for specific customer segments. When the δ_r are time-constant (equal to δ), the value function is a geometric series:

$$val(R_t) = \sum_{R_t}^{r=0} \delta(1+d)^{-r} = \delta\left(\frac{1+d}{d}\right)(1-(1+d)^{-(R_t+1)}). \tag{2.13}$$

In this case, CLV follows an exponential growth curve, increasing at a decreasing rate from δ to an upper asymptote:

$$lim_{R_t \to +\infty}(val(R_t)) = \delta\left(\frac{1+d}{d}\right). \tag{2.14}$$

As the discount rate d approaches zero the value function becomes linear:

$$lim_{d_0 \to +0}(val(R_t)) = \delta(R_t+1). \tag{2.15}$$

Tipically δ_r increase with customer tenure because of increasing revenue growth, operating cost savings, referrals, and a price premium. The form of the functions can be estimated using regression on historical customer cash flows. If the δ_r increases linearly, $\delta_r = \alpha_0 + \alpha_1 r$ for $(\alpha_1 > 0)$, then the value function is:

$$\begin{aligned} val(R_t) &= \sum_{r=0}^{R_t}(\alpha_0+\alpha_1 r)(1+d)^{-r} \\ &= \alpha_0\frac{1}{d}(1+d-(1+d)^{-R_t}) + \alpha_1\sum_{j=1}^{R_t}\sum_{r=j}^{R_t}(1-d)^{-r} \\ &= \frac{1+d}{d}\left(\alpha_0+\frac{\alpha_1}{d}\right) - \frac{(1+d)^{-R_t}}{d}\left(\alpha_0+\frac{\alpha_1(1+d)}{d}+\alpha_1 R_t\right). \end{aligned} \tag{2.16}$$

CLV increases sigmoidally from 0 to an upper asymptote at:

$$lim_{R_t \to +\infty}(val(R_t)) = \frac{1+d}{d}\left(\alpha_0+\frac{\alpha_1}{d}\right). \tag{2.17}$$

As the discount rate d approaches zero, the inflection point moves toward infinity and value function becomes an increasing quadratic curve:

$$lim_{d\to 0}(val(R_t)) = \frac{(2\alpha_0 + R_t\alpha_1)(R_t + 1)}{2}. \tag{2.18}$$

Abrupt falls in the individual cash flows are caused by dropping products or services. Banking customers can close their money market account but keep their checking account. Insurance customers can drop their vehicle policy but keep their homeowners policy. When a company offers several detached products or services, these abrupt changes can be avoided by modelling product churn instead of customer churn. The cash flows are separated by product and separate churn models are built for each product. CLV is the sum of the product-level lifetime values within a customer.
Abrupt rises in the individual cash flows are caused by adding products or services. When the probability of these events is small, incorporating predictive models for add-on buying can cause greater error than assuming no product additions.

2.2.3 Hazard modelling

The churn hazard can be estimated using customer history data extracted from company databases. The outcome is the duration from inception to churn. Customers who are still active at the observation date are censored. The available data is often left-truncated to exclude customers who churned before some date. The data also contains customer covariates such as the product changes, usage, marketing channel, and demographics. Many of the covariates are time-dependent. Hazard models are used to estimate the shape of the hazard function (the time effect) and how the shape is affected by the covariates. Predictive hazard models can be used to score customers. The inputs are the current customer tenure and the current values of the other covariates. The output is the churn hazard at that time the conditional probability of churn.
Plugging-in the maximum likelihood estimate of the hazard gives the maximum likelihood estimate of the mean CLV:

$$\delta_0 + \sum_{r=1}^{\infty} \delta_r (1+d)^{-r} \prod_{j=t}^{t+r-1} (1 - \hat{h}(j|x(j))). \tag{2.19}$$

We cannot compute the formula directly. But this infinite series converges for certain value functions and certain assumptions about the hazard. Assume that after some

time point z the hazard becomes constant at its current value: $h(j) = h(z)$ for $j \leq z$. The time point z could be the maximum tenure in the data or longer if you trust the ability of the hazard model to extrapolate is trusted. Under this assumption, the mean converges for the value function with constant cash flows $\delta_r = \delta$.

$$\begin{aligned}
E(val(R_t)) &= \delta + \delta \sum_{r=1}^{\infty} (1-d)^{-r} \prod_{j=t}^{t+r-1} (1 - \hat{h}(j|x(j))) \\
&= \delta + \delta \sum_{r=1}^{z-t} \left((1-d)^{-r} \prod_{j=t}^{t+r-1} (1 - \hat{h}(j|x(j))) \right) \\
&+ \delta \sum_{r=z-t+1}^{\infty} (1+d)^{-r} \prod_{j=t}^{z-1} (1 - h(j|x(j))) \prod_{j=z}^{t+r-1} (1 - h(z|x(z))) \\
&= \delta + \delta \sum_{r=1}^{z-t} \left((1-d)^{-r} \prod_{j=t}^{t+r-1} (1 - \hat{h}(j|x(j))) \right) \\
&+ \delta (1+d)^{-(z-t)} \frac{1 - h(z|x(z))}{d - h(z|x(z))} \prod_{j=t}^{z-1} (1 - h(j|x(j))).
\end{aligned} \tag{2.20}$$

The formula is not pretty, but all the sums and the products are finite. The value function with linear cash flows is tractable as well. However for many business problems, it is more sensible to base CLV on the restricted residual life:

$$\min(R_t, \gamma + 1), \tag{2.21}$$

the remaining time until churn or an upper bound whichever comes first. This new random variable replaces R_t with an upper bound $\gamma + 1$ whenever R_t exceeds γ. The value of a customer over the next $\gamma + 1$ time units is a function of the restricted residual life:

$$val(min(R_t, \gamma + 1)) = \sum_{r=0}^{min(R_t, \gamma+1)} \delta_r (1+d)^{-r}. \tag{2.22}$$

The mean restricted CLV is the finite sum of a finitely expanding product:

$$E(val(min(R_t, \gamma + 1))) = \delta_0 + \sum_{r=1}^{\gamma+1} \delta_r (1+d)^{-r} \prod_{j=t}^{t+r-1} (1 - h(j|x(j))). \tag{2.23}$$

This is the expected value of a customer over the next $\gamma + 1$ time units. The mean restricted CLV is more sensibly interpreted as the expected long-term value. The upper bound $\gamma + 1$ might represent the limit on a meaningful remaining lifetime. So that customers whose residual life exceeds this time horizon are considered equal. The discount factor causes diminishing returns for longer lifetimes. The upper bound could be chosen as the time when the value function is within c units of the upper

asymptote. For the value function with constant cash flow, choosing γ to be the smallest integer greater than:

$$\frac{ln(\delta/c) - ln(d)}{ln(1+d)}, \tag{2.24}$$

ensures that CLV is within c units of its maximum. The restricted mean CLV can be computed for any given value function. It requires computing the sum of an expanding product, such as:

$$\begin{aligned}
&\delta_0 \\
&+\delta_1(1+d)^{-1}(1-h(t|x(t))) \\
&+\delta_2(1+d)^{-2}(1-h(t|x(t)))(1-h(t+1|x(t+1))) \\
&+... \\
&+\delta_{\gamma+1}(1+d)^{-(\gamma+1)} \times \\
&\times(1-h(t|x(t)))(1-h(t+1|x(t+1)))...(1-h(t+\gamma|x(t+\gamma))).
\end{aligned} \tag{2.25}$$

This is computationally inefficient, because it would involve repeatedly calculating the hazard for the same customer at the same time. A more efficient method is to rearrange sum of the products so that the elements of the product only appear once. The general form can be factored:

$$v_1a_1 + v_2a_0a_1 + v_3a_0a_1a_2 + v_4a_0a_1a_2a_3 = a_0(v_1 + a_1(v_2 + a_2(v_3 + a_3v_4))). \tag{2.26}$$

The a_j represent the elements of the product $1 - h(j)$ and the v_r represent the coefficients of the products $\delta_r(1+d)^{-r}$. On the right-hand-side, each a_j only appears once.

The median CLV can be computed by simply plugging in the median residual life:

$$med(val(R_t)) = val(med(R_t)). \tag{2.27}$$

This convenient identity does not hold with the mean unless the value function is linear. For the value function with constant cash flows the median equals:

$$med(val(R_t)) = \delta\left(\frac{(1+d)}{d}\right)(1-(1+d)^{-(med(R_t)+1)}. \tag{2.28}$$

For the value function with linear cash flows the median equals:

$$\begin{aligned}
med(val(R_t)) &= \frac{(1+d)}{d}\left(\alpha_0 + \frac{\alpha_1}{d}\right) + \\
&-\frac{(1+d)^{-med(R_t)}}{d}\left(\alpha_0 + \frac{\alpha_1(1+d)}{d} + \alpha_1 med(R_t)\right).
\end{aligned} \tag{2.29}$$

Half of all customers of the same tenure and the same values of the other covariates will be worth more than the median, half will not. The probability distribution of CLV is usually not symmetric. If the discount rate d is relatively small, then it is often positively skewed. If d is relatively large, then most of the CLV is close to the upper asymptote and the distribution is negatively skewed. The median CLV is arguably a better measure of centrality than the mean for skewed distributions. The median residual life does not have a closed form but is easy to compute. It is the solution to the equation:

$$\prod_{j=t}^{t+med(R_t)} (1 - h(j|x(j))) = \frac{1}{2}. \tag{2.30}$$

The left hand side is the $Pr(R_t > med(R_t))$ expressed in terms of the hazards. This equation can be solved by iterating from t, updating the product, and stopping when it equals or drops below $\frac{1}{2}$.

2.3 Statistical tenure modelling

From now on we assume that the value is unknown and only the tenure part is to be estimated. This can be done by statistical modelling. A variety of statistical techniques arising from medical survival analysis can be applied to tenure modelling. Tenure prediction models we have developed generate, for a given customer i, a hazard curve or a hazard function, that indicates the probability $h_i(t)$ of churn at a given time t in the future. A hazard curve (see e.g. Hougaard, 1995) can be converted to a survival curve or to a survival function which plots the probability $S_i(t)$ of 'survival' (non-churn) at any time t, given that customer I was 'alive' (active) at time *t-1*, i.e.,

$$S_i(t) = S_i(t-1) \times [1 - h_i(t)], \tag{2.31}$$

with $S_i(1) = 1$. Once a survival curve for a customer i is available, LTV for that specific customer i can be computed as:

$$LTV = \sum_{t=1}^{T} S_i(t) \times v_i(t), \tag{2.32}$$

where $v_i(t)$ is the expected monetary value of customer i at time t (assumed to be known) and T is the maximum time period under consideration.

Survival analysis is concerned with studying the time between entry to a study and

a subsequent event (churn). That is, the times at which events occur are assumed to be realizations of some random processes (see e.g. Klein and Moeschberger, 1997). It follows that T, the event time for some particular individual, is a random variable having a probability distribution. A useful, model-free approach for all random variables is nonparametric, that is, based on cumulative distribution function, $F(T) = P(T \leq t)$. In survival analysis it is more common to work with the survivor function defined by $S(t) = P(T > t) = 1 - F(t)$. If the event of interest is a churn the survivor function gives the probability of surviving beyond t. Because S is a probability we know that it is bounded by 0 and 1 and because T cannot be negative, we know that $S(0) = 1$. Finally, as t gets larger, S never increases (see e.g. Kaplan and Meier, 1958).

A survival function is typically estimated through the methodology of Kaplan Meier (KM). When there are non censored data the KM estimator is just the sample proportion of observations with event times greater than t. The situation is also quite simple in the case of single right censoring, that is, when all the censored cases are censored at the same time c and all the observed event time are less then c. In that case, for all $t \leq c$ the KM estimator is still the sample proportion of observations with events time greater then t. For $t > c$ the estimator is undefined. Things get more complicated when some censoring times are smaller then some event times. In that instance, the observed proportion of cases with event times greater then t can be biased downward because cases that are censored before t may, in fact, have 'died' before t without our knowledge. A possible solution is as follows. Suppose there are K distinct event times, $t_1 < t_2 < ... < t_k$. At each time t_j there are n_j individuals who are said to be at risk of an event. At risk means they have not experienced an event not have they been censored prior to time t_j. If any cases are censored at exactly t_j, there are also considered to be at risk at t_j. Let d_j be the number of individuals who die at time t_j. The KM estimator is defined as:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{n_j}\right], t_1 \leq t \leq t_k, \tag{2.33}$$

This formula says that, for a given time t, all the event times that are less than or equal to t are taken. For each of those event times, the quantity in brackets, which can be interpreted as the conditional probability of surviving to time t_{j+1}, given that one has survived to time t_j, is computed and finally all of these survival probabilities

are multiplied together.

Often the objective is to compare survivor functions for different subgroups in a sample (clusters, regions). If the survivor function for one group is always higher than the survivor function for another group, then the first group clearly lives longer than the second group. Two or more survival functions can be formally compared by means of Scheffe tests.

When variables are continuous, another common way of describing their probability distributions is the probability density function, defined as:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \tag{2.34}$$

From the previous equation note that the probability density function is just the derivative or slope of the cumulative distribution function. For continuous survival data, the hazard function is actually more popular than the probability density function as a way of describing distributions.

2.3.1 Classical Cox Model

A hazard function is defined by:

$$h(t) = lim_{\epsilon_t \to 0} \frac{Pr(t \leq T \leq t + \epsilon_t | T \geq t)}{\epsilon_t}. \tag{2.35}$$

The hazard function depends in general on both time and a set of covariates, some of which may be time dependent. The proportional hazards model Cox (1972), separates these components by specifying that the hazard at time t for an individual whose covariate vector is x is given by:

$$h(t|x) = h_0(t) exp\left\{G(x, \beta)\right\}, \tag{2.36}$$

where $h_0(t)$ is called the baseline hazard function and β is a vector of regression coefficients. This model implies that the ratio of the hazards for two individuals is constant over time, provided that the covariates do not change. It is conventional to assume that the effect on the covariates is multiplicative, leading to the hazard function:

$$h(t|x) = h_0(t) exp\left\{x\acute{}\beta\right\}, \tag{2.37}$$

where $\eta = x\acute{}\beta$ is called the linear predictor.

The model in equation (2.37) implies that the ratio of hazards for two individuals

depends on the difference between their linear predictors at any time. Suppose that there are n subjects and that associated with the i^{th} individual is a survival time t_i and a fixed censoring time c_i. The t_i times are assumed to be independent and identically distributed with density $f(t)$ and survival function $S(t)$. The exact survival time t_i of an individual will be observed only if $t_i \leq c_i$. The data in this framework can be represented by the n pairs of the random variables (y_i, ν_i), where

$$y_i = \min(t_i, c_i), \tag{2.38}$$

and ν_i is 1 if $t_i \leq c_i$ and 0 otherwise.

The likelihood function in this standard case for $(\beta, h_0(.))$ for a set of right censored data on n subjects is given by

$$\begin{aligned} L(\beta, h_0(t)|D) &\alpha \prod_{i=1}^{n} [h_0(y_i)exp(\eta_i)]^{\nu_i} \left(S_0(y_i)^{exp(\eta_i)}\right) \\ &= \prod_{i=1}^{n} [h_0(y_i)exp(\eta_i)]^{\nu_i} \, exp\left\{-\sum_{i=1}^{n} exp(\eta_i)H_o(y_i)\right\}, \end{aligned} \tag{2.39}$$

where $D = (n, y, X, \nu)$, $y = (y_1, ..., y_n)'$, $\nu = (\nu_1, ..., \nu_n)'$, $\eta_i = x_i'\beta$ is the linear predictor for subject i, x_i is a vector of covariates for subject i, X is a matrix of covariates and $S_0(t)$ is the baseline survivor function, which is related to $h_0(.)$ by:

$$S_0(t) = exp\left(-\int_t^0 h_0(u)du\right) = exp(-H_0(t)). \tag{2.40}$$

The Cox version of proportional hazards model is semi-parametric in the sense that the baseline hazard function $h_0(t)$ is not modelled as a parametric function of t. In Cox's development of the partial likelihood (Cox, 1972), $h_0(t)$ is allowed to take on arbitrary values since it does not enter into the estimating equations for the model parameters.

Suppose that data are available on n individuals, and assume from these that we have d distinct events times and $n - d$ right censored survival times. Denote the ordered distinct survival times by $y_1, ..., y_d$, so that y_j is the j^{th} survival time. The set of individuals who are at risk at time y_j will be denoted by R_j, that is the set of individuals who are event-free and uncensored at a time just prior to y_j. The Cox partial likelihood for β is defined by:

$$PL(\beta|D) = \prod_{j=1}^{d} \frac{exp(x_j'\beta)}{\sum_{l \in R_j} exp(x_l'\beta)}, \tag{2.41}$$

where the summation in the denominator express the sum of the values of $exp(x'_j\beta)$ over all individuals who are at risk at time y_j. This likelihood depends only on the ranking of the events times, since this determines the risk set at each event time. Consequently inferences about β depend only on the rank order of the survival times.

2.3.2 Criticism on the classical Cox Model

A very crucial aspect of causal models in survival analysis is the preliminary stage, in which a set of explanatory variables must be properly chosen and designed, usually among a very large number of alternatives. This part of the analysis is typically accomplished with the help of descriptive tools, such as plots of the observed hazard rates at the covariate values. However, it is often the case that such tools are not sufficiently informative. As a consequence, a large number of variables are included as predictors and a model selection procedure needs to be run in order to find a parsimonious linear combination. Our claim is that classical Cox proportional hazard models may not be the best strategy for Customer Lifetime Value modelling. Some criticisms are:

- If repeated event occurs, as in our case, a different model structure (e.g. based on counting processes) should be adopted.

- Cox model assumes that every subject experiences an event at most once, and that the event times are independent. In our context, a subject can experience multiple events (e.g. a churn event in different times and locations), possibly with dependencies among the event times of the same individual. Modelling multiple event time data requires a different approach. An example of modelling multiple event time data was given by Gail, Santer and Brown (1980) with an application on mammary tumor.

- When many explanatory variables, possibly correlated, are specified, the *efficiency* of Cox's model selection and estimation becomes heavily dependent on the number of available observations. Variable selection is thus needed in a model selection step. However classical model selection chooses on model and then inferences on quantities of interest, such as $\lambda(t|z)$ are then made *conditionally* upon the selected model. Consequently, model uncertainty is not taken into account and, thus, inference may be seriously biased.

- It may be difficult, particularly in observational studies, to have *complete* information on all relevant covariates. Furthermore, random effects, expressing accident proneness or *frailties* may affect inferences on fixed effects.

In this dissertation we shall show how to improve the classical Cox model, tackling some of the above criticisms for the specific problem at hand. More precisely:

- We shall consider a point process framework to model repeated events;
- We shall consider Bayesian variable selection and Bayesian model averaging to correctly take model uncertainty into account. We shall also propose a new method for models grouping to improve the efficiency of model averaging;
- We shall introduce a multilevel multivariate survival model via stratification, to take frailties into account;

2.4 Survival analysis in the point processes framework

Several authors have discussed Bayesian Inference for censored survival data with an integrated baseline hazard function to be estimated non-parametrically: Kalbfleisch (1978), Kalbfleisch and Prentice (1980). In particular, Clayton (1994) formulates the Cox Model using the counting process notation introduced by Andersen and Gill (1982) and discusses estimation of the baseline hazard and regression parameters using a Bayesian approach based on Markov Chain Monte Carlo. Although his approach may appear somewhat contrived, it forms the basis for extensions to random effects frailty models, time-dependent covariates, smoothed hazards, multiple events and so on.

Here we follow Clayton's guidelines and propose a methodology based on counting processes. In particular the counting process we present is characterized by a dynamic process (intensity), and a special pattern of incompleteness of observations (right-censoring or left-truncation). This characterization is an application of the well known Doob-Meyer decomposition theorem. Having defined the intensity process, one is interested in estimation of its parameters.

Inferential procedures in this framework were first presented in Aalen Hoem (1980), and turned out to be very fruitful. For further developments, see Anderson et al.

(1993). A counting process is a stochastic process $\{N(t) : t \geq 0\}$ adapted to a self-exciting filtration $Im_t : t \geq 0$ with $N(0) = 0$ and $N(t) < \infty$ a.s., whose paths are with probability one right-continuous, piecewise constant, with only jump discontinuities, with jumps of size one.

For the derivation of the likelihood function we follow a well developed theory leading to a Poisson type of likelihood (see e.g. Andersen et. al. (1993), or Fleming and Harrington (1991)). The argumentation is based on Jacod's Formula for the likelihood ratio.

We consider the analysis of multiple event data where there are n groups and the i^{th} cluster has m_i individuals associated with an unobserved frailty w_i, $1 \leq i \leq n$. The j^{th} individual in the i^{th} cluster, is associated with the fixed covariate vector x_{ij}. Such individuals are assigned as belonging to a specific cluster because they are related somehow, say by family association or graphical location. Conditional on frailties w_i, the complete survival times are assumed to be independent. For convenience we suppress for the time being the subscript indexing individuals and clusters and consider the model:

$$h(t|w, x) = w[h_0(t) + h_1(t|x)], t \leq 0, \tag{2.42}$$

where $h(t|w, x)$ represents a hazard function that has been modified by the inclusion of a frailty.

The frailty random variable, w, is assumed to be independent of t and x for all clusters with some parametric distribution with unit mean, usually Gamma (Clayton, 1991), which the unknown variance of w, say η, that quantifies the amount of heterogeneity among individuals. In other words we assume that $w_i|\eta \sim Ga(\eta^{-1}, \eta^{-1})$, that is, given η, w_i, $i = 1, ..., n$, is modelled as Gamma distribution with scale parameter η^{-1} and shape parameter η^{-1}. It is important to keep in mind the interpretation of w in the hazard function $h(t|w, x)$. The frailty random variable w measures the random sensitivity of the $i - th$ cluster to the event of interest after taking into account the effect of the covariate. More precisely, as the frailty ω acts as a multiplicative factor in (2.42) when the value of the frailty is greater than the mean, the individual has a larger then average hazard and is said to be more "frail" and vice versa. That is why for finite mean frailty we need to assume unit mean (to assure identifiably) and we need to assume that the frailty distribution of individuals at

different covariate levels share the same mean but may have different variability.

The non parametric part of the model, $h_0(t)$, is assumed to be a piecewise linear hazard. An ordinary piecewise constant hazard (see e.g. Gamerman, 1991), which is an example of a semi-parametric hazard specification, has the advantage that it is a simple way to get a flexible hazard function, with simple estimation. On the other hand, it has a major disadvantage, because the hazard is not continuous as a function of time, as there are jumps at the interval end points. In order to avoid such discontinuities we can use an ordinary piecewise exponential model (Gamerman,1991).

To construct such model, we first split the time axis into intervals $0 = a_0 < a_1 < ... < a_g$, where g is the number of intervals of observations times, $a_g > t_{ij}$ for all $i = 1, 2, ..., n$ and $j = 1, 2, ..., m_i$. The hazard in the interval $I_k = (a_{k-1}, a_k]$ is defined by $\lambda_{00} + (t - a_{k-1})\lambda_{0k} I(a_{k-1} < t)$.

Therefore, the hazard function is obtained as:

$$h_0(t) = \lambda_{00} + \sum_{k=1}^{g} (t - a_{k-1})\lambda_{0k} I(a_{k-1} < t), \tag{2.43}$$

where the function I is the indicator function whose value is one if the argument is true and zero otherwise. Note that $h_0(t)$ has one parameter more than the usual piecewise constant hazard model. Concerning the third part of the model, we assume that the covariates are time independent and that the hazard is a constant function of t. This leads to an exponential distribution with θ parameter so that $h(t|x) = \frac{1}{\theta}$, if $t > 0$.

In general, the survival function is related to the hazard function through the expression $S(t) = exp[-H(t)]$, where $H(T) = \int_0^t h(u)du = -ln[S(t)]$. Given the relationship between the hazard and the survival function, modified by the inclusion of a frailty component, specified by the parameters $w = (w_1, ..., w_n)$, $\lambda_0 = (\lambda_{00}, ..., \lambda_{0g})$ and θ is:

$$S(t|w, \lambda_0, \theta) = [S_0(t|\lambda_0) S_1(t|\theta)]^w, \tag{2.44}$$

where $S_0(t)$ and $S_1(t)$ are, respectively the survival functions related to the piecewise exponential baseline hazard function and exponential hazard function, i.e.,

$$\begin{aligned} S_0(t|\lambda) &= exp[-\int_0^t h_0(u)du] = exp[-\sum_{k=0}^{g} c_k(t)\lambda_{0k}], \\ S_1(t|\theta) &= \int_0^\infty \frac{1}{\theta} exp\left(-\frac{u}{\theta}\right) du = exp\left(-\frac{t}{\theta}\right), \end{aligned} \tag{2.45}$$

where $c_k(t)$ are positive constants. Therefore, the density function takes the form:

$$\begin{aligned} f(t|w,\lambda_0,\theta) &= h(t|w,\lambda_0,\theta)S(t|w,\lambda_0,\theta), \\ &= w[h_0(t|\lambda_0) + h_1(t|\theta)][S_0(t|\lambda_0)S_1(t|\theta)]^w. \end{aligned} \tag{2.46}$$

We assume that the parameter θ is specific for each individual in the population, but related to the covariate x through a probabilistic model.

In order to facilitate the implementation, it is convenient to assume that will be modelled as a Normal distribution with mean βx: a linear combination of the effects covariates such that $\beta = (\beta_1, ..., \beta_p)$ and x be the $N \times p$ matrix with rows $x_1, ..., x_N$; and variance σ_θ^2. The hyperparameters β and σ_θ^2 are unknown parameters common to all individuals in the population.

Note that the described model allows two sources of variability: one due to covariates and one due to frailties. Indeed we can find two different individuals who have the same covariate vector, but their hazard functions are not necessarily identical, because of the frailty effect between cluster heterogeneity, described by the parameter of the vector parameter ω_i, $i = 1, ..., n$.

In other words, $\theta_i = \sum_{l=1}^{p} \beta_l x_{il}$ denote the expected value of death for j^{th} individual in the cluster i.

A concise hierarchical representation of the model enables us to implement the MCMC methodology, that allows the Bayesian analysis of the problem. It is as follows

Suppose that the y^{th} individual in the i^{th} cluster survival time T_{ij} is an absolutely continuous random variable conditionally independent of a right censoring time Z_{ij} given the covariates x_{ij} and frailty w_i.

Let $Vij = min(T_{ij}, Z_{ij})$ and $\delta_{ij} = I(T_{ij} \leq Z_{ij})$ denote the time to the end-point event and the indicator for the event of interest to take place, respectively. Suppose that $(V_{ij}, \delta_{ij}, x_{ij}, w_i)$ are i.i.d, for $i = 1, ..., n$ and $j = 1, ...m_i$, and the conditional hazard function of T_{ij} given x_{ij} and w_i satisfies the hazard model described by (2.42). For subject j in cluster i, let $N_{ij}(t) = 1$ if $\delta_{ij} = 1$ is in the interval $[0, t]$ and $N_{ij}(t) = 0$ otherwise, and let $Y_{ij}(t) = 1$ if the subject is still exposed to risk at time t and $Y_{ij}(t) = 0$ otherwise.

Hence, we have a set of $N = \sum_{n}^{i=1} m_i$ subjects such that the counting process $\{N_{ij}(t); t \geq 0\}$ for the j^{th} subject in i^{th} cluster set, records the number of observed

events up to time t. Letting $dN_{ij}(t)$ denote the increment of $N_{ij}(t)$ over the small interval $[t, t+dt]$, the likelihood of the data conditional on w_i is then proportional to:

$$L(\lambda_0, \omega, \theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} \left(\prod_{t \geq 0} Y_{ij}(t) w_i \left[h_0(t|\lambda_0) + h_1(t|\theta) \right]^{dN_{ij}(t)} \right)$$
$$\times exp\left(- \int_{t \geq 0} Y_{ij}(t) w_i \left[h_0(t|\lambda_0) + h_1(t|\theta) \right] \right), \tag{2.47}$$

Since we allow each $N_{ij}(t)$ to take at most one jump for each subject, note that $dN_{ij}(t)$ contribute to the likelihood in the same manner as independent Poisson random variables even though $dN_{ij}(t) \leq 1$ for all i, j and t.

Suppose that the time axis is partitioned into $g+1$ disjoint intervals $I_1, ..., Ig+1$ where $I_k = [a_{k-1}, a_k)$ for $K = 1, 2, ..., g+1$, with $a_0 = 0$ and $a_{g+1} = \infty$. In the K^{th} interval, given w_i, the j^{th} subject in the i-th cluster has an hazard equal to: $w_i \{h_0(t_{ij}|\lambda_{0k}) + h_1(t_{ij}|\theta_{ij})\}$, $K = 1, ..., g_{ij}$ where g_{ij} denotes the number of partitions of the time interval for the j^{th} subject in the i^{th} group.

Given the complete data (T, w), where $T = \{t_{ij} : i = 1, ..., n_i; j = 1, ..., m_i\}, w = (w_1, ..., w_n)$, the likelihood can be re-expressed as:

$$L(\lambda_0, \omega, \theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} \prod_{k=1}^{g_{ij}} \prod_{t \in (a_{k-1}, a_k)} \left[Y_{ij}(t) w_i \left[h_0(t|\lambda_0) + h_1(t|\theta) \right]^{dN_{ijk}} \right]$$
$$\times exp\left(- \int_{t \geq 0} Y_{ij}(t) w_i \left[h_0(t|\lambda_0) + h_1(t|\theta) \right] \right), \tag{2.48}$$

where dN_{ijk} is the change in the count function for the j^{th} subject in the i^{th} group in the interval k. Under the assumption that the risk occurred in the interval I_k is small, i.e.,

$$\int_{a_k}^{a_{k-1}} Y_{ij}(t) \left[h_0(t|\lambda_0) + h_1(t|\theta) \right] dt \approx 0 \tag{2.49}$$

for all i,j,k, the likelihood contribution across this interval for individuals at risk is approximately:

$$\left\{ w_i \left[dH_{0k} + \frac{1}{\theta_{ij}} (a_k - a_{k-1}) \right] \right\}^{dN_{ijk}} \times exp\left(-w_i \left[dH_{0k} + \frac{1}{\theta_{ij}} (a_k - a_{k-1}) \right] \right), \tag{2.50}$$

where $dH_{0k} = \int_{ak-1}^{a_k} h_0(t) dt$ is the usual cumulative baseline intensity function for the k^{th} interval.

Notice again that the likelihood is essentially Poisson in form, reflecting the fact that the likelihood may be thought of as a generated by independent contributions of

many data 'atoms' each concerned with observation of an individual over a very short interval during which the intensity may be regarded constant and approximately zero (for a review of this point, see e.g. Clayton, 1994). Substituting (2.49) into (2.47), can express the likelihood as:

$$L(\lambda_0, \omega, \theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{m_i} \prod_{Y_{ijk}=1} \left\{ w_i \left[dH_{0k} + \frac{1}{\theta_{ij}} (a_k - a_{k-1}) \right] \right\}^{dN_{ijk}}$$
$$\times exp\left(-w_i \left[dH_{0k} + \frac{1}{\theta_{ij}} (a_k - a_{k-1}) \right] \right), \tag{2.51}$$

where $Y_{ijk} = 1$ if the j^{th} subject in the i^{th} group is exposed to risk at time $t \in (a_{k-1}, a_k]$, and $Y_{ijk} = 0$.

2.5 Bayesian survival analysis

Several authors have addressed survival analysis models from a Bayesian viewpoint. There are may references, see e.g. Anderson, Borgan, Gill and Keiding (1993), Berzuini and Clayton (1994), Ferguson and Phadia (1979), Fleming and Harrington (1991), Giudici, Mezzetti and Muliere (2003), Hastie and Tibshirani (1990), Ibrahim, Chen and Sinha (2001a), Laud Smith and Damien(1996), Sinha et al. (2003), Walker and Mallick (1996). One common problem present in Bayesian modelling of survival data is the presence of a large numbers of alternative models that verify the assumptions. In practice, one usually needs to address two problems, that is, model adequacy and model comparison.
Here we shall focus on these issues.

2.5.1 Model adequacy

Model adequacy is the problem of selecting the 'right form' in the model. Classical model adequacy is typically checked with cross-validation criteria (see e.g. Giudici, 2003, Hastie Tibshirani and Friedman 2001). The literature for model adequacy in a Bayesian framework does not seem to be rich. A formal Bayesian model adequacy criterion (as in Box, 1980) proposes that the marginal predictive density is to be evaluated at the actual times of observations. Large values of the density support the model; small values do not. A specific proposal, is as follows. Let y_{obs} be the observed data, θ be the vector of unknown parameters in the model. We

assume that we have draws $\theta_1, ..., \theta_n$ from the posterior distribution, possibly using Markov Chain simulations. We now simulate N hypothetical replications of the data $y_1^{new}, ..., y_N^{new}$, where y_i^{new} is drawn from the predictive distribution of y_{obs}. Thus y^{new} has distribution:

$$P\left(y^{new}|y_{obs}\right) = \int P(y^{new}|\theta)P\left(\theta|y_{obs}\right)d\theta, \tag{2.52}$$

One can then compare the actual data to the predicted values by choosing a discrepancy variable test statistic, which will have a large value if the data y_{obs} are in conflict with the model.

Unfortunately, such checking procedures seem to be technically very difficult, or even unfeasible, for most survival data analysis problems. A more feasible approach is based on the calculation of Bayes factors.

Let w_i be the prior probability of model M_i, $i = 1, 2$, and $f(y|M_i)$ be the predictive distribution under model M_i, i.e.

$$f(y|M_i) = \int f(y|\theta_i, M_i)\pi(\theta_i|M_i)d\theta_i \tag{2.53}$$

If y_{obs} denotes the observed data then the Bayes factor of model M_1 w.r.t. model M_2 is defined by:

$$BF = \frac{f(y_{obs}|M_1)}{f(y_{obs}|M_2)} \tag{2.54}$$

Schwarz (1978) derived the *Bayesian Information Criterion* (or BIC) as a large sample approximation to twice the logarithm of the Bayes factor. For a model M_j parameterized by an $m_j - dimensional$ vector θ_j,

$$BIC = -2\left\{l_j(\hat{\theta}_j) - l_0(\hat{\theta}_0)\right\} + (m_j - m_0)\log(n), \tag{2.55}$$

where $l_j(\hat{\theta}_j)$ and $l_0(\hat{\theta}_0)$ are the maximized likelihoods under model M_j and a reference model M_0, whose parameter space has dimension m_0 and n is the sample size.

With nested models, BIC equals the standard likelihood ratio test statistic plus a complexity penalty which depends on the degrees of freedom of the test of M_0 against M_j. BIC provides an approximation to the Bayes Factor which can be computed from the output of standard statistical software packages (see e.g. Kass and Raftery 1995, Raftery 1995). The derivation of BIC involves a Laplace approximation to the Bayes Factor, and ignores terms of constant order, so that the difference between BIC and twice the log Bayes Factor does not vanish asymptotically in general, although

it becomes inconsequential in large samples. Kass and Raftery (1995) derive BIC as an approximation to twice the difference in log integrated likelihoods, so that the difference in BIC between two models approximates twice the logarithm of the Bayes factor. More formally:

$$\frac{2\log(BF) - BIC}{2\log(BF)} \to 0; \tag{2.56}$$

however,

$$2\log(BF) - BIC \neq 0, \tag{2.57}$$

The last equation implies that, for general priors on the parameters, $2\log(BF) - BIC$ has a no-vanishing asymptotic error of constant order, i.e. of order $O(1)$.This $O(1)$ error suggests that the BIC approximation is somewhat crude, and may perform poorly for small samples.

Kass and Wasserman (1995) show that with nested models, under a particular prior on the parameters, the constant order asymptotic error disappears, and they argue that this prior can reasonably be used for inference purposes. Following the notation of their paper, let $Y = (y_1, ..., y_n)$ be i.i.d. observations from a family parameterized by (θ, ψ), with $dim(\theta, \psi) = m$ and $dim(\theta) = m_0$. If the goal is to test $H_0 : \psi = \psi_0$ against $H_1 : \psi \in \Re^{m-m_0}$ using the Bayes factor:

$$BF = \frac{P(Y|H_0)}{P(Y|H_1)}, \tag{2.58}$$

the Bayesian Information Criterion (BIC) for testing H_0 vs. H_1 is:

$$BIC = -2\left\{l_1(\hat{\theta}, \hat{\psi}) - l_0(\hat{\theta}_0)\right\} + (m_j - m_0)\log(n). \tag{2.59}$$

Let $I(\theta, \psi)$ be the $m \times m$ Fisher information of (θ, ψ) associated with the full model, let $I_{\theta\psi}(\theta, \psi_0)$ denote the information matrix $\left(-E(\frac{\partial^2 l(\theta,\psi)}{\partial\theta\partial\psi})\right)$, evaluated at (θ, ψ_0), and let $\pi_\psi(\psi)$ be the marginal prior density of ψ under H_1.

The main results of Kass and Wasserman (1995) is as follows. If the following conditions hold:

- the parameters are *null orthogonal*, that is, $I_{\theta\psi}(\theta, \psi_0) = 0$ for all θ,
- the MLE $\hat{\psi}$ satisfies $\hat{\psi} - \psi_0 = O_p(n^{-\frac{1}{2}})$, and
- $-\frac{1}{n}D^2 l(\hat{\theta}, \hat{\psi}) - I(\theta, \psi) = O_p(n^{-\frac{1}{2}})$,

then

$$2\log BF = BIC - 2\log(2\pi)^{\frac{(m-m_0)}{2}}|I_{\psi\psi}(\hat{\theta},\psi_0)|^{-\frac{1}{2}}\pi_\psi(\hat{\psi}) + O_p(n^{-\frac{1}{2}}). \tag{2.60}$$

In addition if $\pi_\psi(\psi)$ is a standard multivariate normal density with location ψ_0 and variance matrix $|I_{\psi\psi}(\theta,\psi_0)|^{-1}$ the asymptotic error of constant order will vanish, leaving

$$2\log(BF) = BIC + O_p(n^{-\frac{1}{2}}), \tag{2.61}$$

If the prior on ψ is not of this form, then this error term gives the constant order asymptotic error in BIC as an approximation to twice the log Bayes factor.
This result has an important implication. BIC is a Bayesian procedure which does not require the specification of a prior, but it approximates a Bayes factor which is based on a particular prior for the parameter of interest. Therefore, when using BIC to compare models, the Kass-Wasserman result defines an *implicit* prior which BIC uses. This prior, which we call the *overall unit information prior*, is appealing: it is a normal distribution centred around ψ_0 with the amount of information in the prior equal to the average amount of information in one observation. Since the prior is based on only one observation, it is a vague but proper prior.

2.5.2 Model comparison

Model comparison is required for a variety of activities, including variable selection in regression, determination of the number of components in a mixture model or the choice of a parametric family. As with frequentist analogues, Bayesian model comparison will not inform about which model is "true", but rather about the preference for the models given the data.
These preferences can be used to choose a "representative" best model or improve estimation via model averaging, in which expected values obtained from different models are weighted by their corresponding posterior probabilities.
In the Bayesian framework, common methods for model comparison are based on the following: *separate estimation* including posterior predictive distributions, Bayes Factors and approximations such as the Bayesian Information Criterion (BIC) and deviance information criterion (DIC); *comparative estimation* including distance measures such as entropy distance or Kullback-Leibler divergence; and *simultaneous estimation*, based on model averaging and computationally intensive MCMC

approaches. While in previous chapter we have focused on separate estimation, we now focus on simultaneous, that is, model averaging procedures.

The classical survival data analysis approach is to select a set of predictors or risk factors and make inferences on the set of predictors using this single model. A serious shortcoming this approach is the dependence of the inferences on the set of predictors selected for inclusion in the model. Bayesian model averaging allows for the incorporation of model uncertainty into inference. The basic idea of Bayesian model averaging is to make inferences based on a weighted average over the model space. This approach accounts for model uncertainty in both predictions and parameter estimates. The resulting estimates of uncertainty incorporate model uncertainty and thus may be better reflect the true uncertainty in the estimates.

We base our approach on the Bayesian model averaging approach suggested by Hoeting, Madigan, Raftery and Volinsky, (1999).

Let $M = (M_1, ..., M_k)$ be the set of models under consideration. A model may be defined by a variety of attributes such as the subset of explanatory variables in the model or the form of the error variance. If Δ is a quantity of interest, such as a future observable or a model parameter, then the posterior distribution of Δ given data Z is:

$$p(\Delta|Z) = \sum_{k=1}^{K} p(\Delta|Z, M_k)p(M_k|Z); \tag{2.62}$$

this is an average of the posterior predictive distribution for Δ under each of the models considered, weighted by the corresponding posterior model probability. The posterior probability for a model M_k is given by

$$p(M_k|Z) = \frac{p(Z|M_k)p(M_k)}{\sum_{l=1}^{K} p(Z|M_l)p(M_l)}, \tag{2.63}$$

where

$$p(Z|M_k) = \int ... \int p(Z|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{2.64}$$

is the integrated likelihood of model M_k, θ_k is the vector of parameters of model M_k, $p(\theta_k|M_k)$ is the prior density of the parameters under model M_k, $p(Z|\theta_k, M_k)$ is the likelihood, and $p(M_k)$ is the prior probability that M_k is the true model. All probabilities are implicitly conditional on M, the set of all models being considered. Parameters estimates and other quantities of interest are provided via straightforward application of the principles described above. For example, the Bayesian model

averaging (BMA) estimate of parameter θ is

$$\hat{\theta}_{BMA} = \sum_{k=1}^{K} \hat{\theta}_k p(M_k|Z), \tag{2.65}$$

where $\hat{\theta}_k$ denotes the posterior mean for model K. Variances of these estimates and other quantities are also available (e.g. Hoeting et al. 1999a).

There are many challenges involved in the implementation of Bayesian model averaging, including the computation for a very large number of models, the evaluation of the integrals implicit in $p(Z|M_k)$ which do not typically exist in closed form, and the specification of the prior model probabilities $p(M_k)$.

A number of researchers have considered the problem of managing the summation in $p(\Delta|Z)$ for a large number of models. A popular approach is to explore the space of models stochastically via Markov Chain Monte Carlo (e.g. George and McCulloch, 1997 and Raftery, Madigan and Hoeting, 1997). Clyde (1999) shows that many of these approaches are a special case of reversible jump MCMC algorithms (Green, 1995). Godsill (2001) and Brooks, Giudici, Roberts (2003) proposes a composite representation for model uncertainty problems which includes many of these Markov Chain Monte Carlo approaches as special cases. Hoeting, Madigan, Raftery, Volinsky (1999) discuss the historical development of BMA, provide additional description of the challenges of carrying out BMA, and describe some solutions to these problems for a variety of model classes. Articles focusing on Bayesian approaches to model comparison in the context of survival analysis include: Sinha, Chen and Ghosh (1999), Ibrahim, Chen and Sinha (2001b), Sahu, Dey, Aslanidou and Sinha (1997), Chen, Harrington and Ibrahim (2001).

Chapter 3

Bayesian lifetime value models

In this chapter we now propose two approaches to Bayesian inference for lifetime value models. A first proposal is a two-step approach, made up of two components:

1. A Bayesian variable selection approach (BVSA) to select the best predictor variables (features).

2. A Bayesian survival model (BSM) that, conditionally on the chosen features, draws inference on the survival parameters.

The second proposal is a one-step approach, which combines (1) and (2) in a Bayesian model averaging procedure.

3.1 A Bayesian variable selection approach

Before to illustrate our proposal for variable selection to estimate Customer lifetime value, we report here the most recent references. We report in particular the most significant results for Variable selection in multi-events survival analysis. Variable selection for multivariate failure time data has been analysed by Fan et al. (2006) based on a penalized pseudo-partial likelihood method. Fan and Li (2005) gives some methods to extract variable selection for parametric models via no-concave penalized likelihood. It has been shown there that the resulting procedures perform as well as if the subset of significant variables were known in advance. Such a property is called "oracle property". Tibshirani (1997) proposed the LASSO methodology. Dunson (2005) proposes a semiparametric Bayesian approach for inference on un

unknown regression function, $f(x)$ characterizing the relationship between a continuous predictor X and a count response variable Y adjusting for covariates, Z.
Giudici, Mezzetti and Muliere (2003) proposed a nonparametric variable selection approach for survival analysis. In general field of variable selection, we remark that Casella and Moreno (2006) proposed a novel fully automatic Bayesian procedure for variable selection in normal regression models, Cai, Fan, Jiang and Zhou (2006) propose a partial linear regression for multivariate survival data and improved Bayesian model averaging with selection of covariates and Gutierrez Pena (2005) for Bayesian methods for categorical data. Finally to improve the results we cite Walker and Gutierrez-Pena for the proposal in Bayesian Parametric inference in a nonparametric framework. To design adaptive mini-Max local linear regression for longitudinal and clustered data we cite Chen, Fan Jin (2006). To improve variable selection with local partial likelihood estimation for life time data, see e.g. Fan, Lin and Zhou (2006). Cai, Fan, Zhou H. and Zhou Y. (2007) we give an idea for Marginal hazard models with varying coefficients for multivariate failure time data.
Our proposal: we follow a novel approach for feature selection and model selection. In particular, we first propose a two step model approach for Lifetime Value estimation. Then we present a new method for Bayesian feature selection based on an One step Lifetime value approach.
To esemplify our variable selection methodology, we shall assume an exponential survival time, such that, for $i = 1, \ldots, n$: $\lambda_i(t) = \lambda_i$. It can then be shown that, given the observed evidence $\underline{y} = (y_1, \ldots, y_n)$, the likelihood of $\underline{\lambda} = (\lambda_1, \ldots, \lambda_n)$ is:

$$L(\underline{\lambda}) = \prod_{i \in \mathcal{U}} \lambda_i \exp\{-\sum_{i=1}^{n} \lambda_i t_i\}, \tag{3.1}$$

where $\mathcal{U} = \{i : \delta_i = 1\}$ are the uncensored subjects. Let now g indicate a *partition* of the index set $\mathcal{I} = \{1, \ldots, n\}$, with d_g subsets $S_k(g)$, for $k = 1, \ldots, d_g$. Clearly, given the correspondence between $\mathcal{I}, \underline{y}$ and $\underline{\lambda}$, g also defines a partition of the data and of the hazard functions. Notice that the likelihood in (**??**) assumes all λ_i to be distinct and, thus, is in fact conditional on the *independence* partition $g_{ind} = \{\{1\}, \{2\}, \ldots, \{n\}\}$, containing $d_g = n$ separate subsets S_i, each with $n(S_i) = 1$ observations. For this reason, it can be indicated by $L(\underline{\lambda}|g_{ind})$.
A different likelihood arises when all hazards can be retained equal to a common rate, say μ. This situation occurs when *no* covariate or frailty affect the survival times

and corresponds to consider all data to be *exchangeable*. The resulting likelihood can be seen as conditional on the partition $g_{exc} = \{1, \ldots, n\}$, containing a single subset S_1 (with $n(S_1) = n$):

$$L(\mu|g_{exc}) = \mu^d exp\{-\mu V\}, \tag{3.2}$$

with $d = \sum_{i=1}^{n} \delta_i$ the total number of failures and $V = \sum_{i=1}^{n} t_i$ the overall time at risk.

Apart from the above situations, which can be regarded as somewhat extreme, survival analysis is typically concerned with a plurality of effects which may induce dependencies among survival times. Such effects may be either observable (possibly with some missing values) or unobservable. In any case, when relevant, they *induce* a partition of the observations, by associating different hazards to individuals having the same level of the factor. In our approach, we shall entertain several partition structures, each induced by the levels of a potential prognostic factor. This amounts to consider a collection of alternative *partial exchangeability structures* for the survival times. Our model consists of two parts: a *likelihood* specification and a *hierarchical prior* distribution on the partition structure as well as on the corresponding set of hazards. Conditionally on a *general* partition g, let $\lambda_i = \mu_k$, $\forall i \in S_k(g)$. Consequently, the likelihood of the hazards $\underline{\mu} = (\mu_1, \ldots, \mu_{d_g})$ is the following:

$$L(\underline{\mu}|g) = \prod_{k=1}^{d_g} \mu_k^{d_k} \exp\{-\mu_k V_k\}, \tag{3.3}$$

where, for $k = 1, \ldots, d_g$: μ_k, $d_k = \sum_{i \in S_k(g)} \delta_i$ and $V_k = \sum_{i \in S_k(g)} t_i$ are the hazard, death and risk set of the k-th partition subset. On the other hand, the prior specification requires the definition of a class of possible partitions $\mathcal{G} = \{1, \ldots, G\}$.

Once $\mathcal{G}$ is specified, it is necessary to assign a probability distribution on both $\underline{\lambda}|g \in \mathcal{R}^{d_g}$ and $g \in \mathcal{G}$. Specifically, conditionally on a partition g we shall take, for $k = 1, \ldots, d_g$ and $\forall i \in S_k(g)$:

$$\mu_k \sim Gamma(r_k m_k, r_k), \tag{3.4}$$

with m_k and r_k known positive constants. Finally, a simple probability function on $\mathcal{G}$ would take $p(g)$ to be uniformly spread among partitions, i.e. $p(g) = G^{-1}$.

Our first aim is to evaluate the importance of each prognostic factor. This can be achieved calculating, given the observed evidence $\underline{y}$, the posterior probability of each

partition, $p(g|\underline{y})$. Following (**??**) and (**??**) it can be shown that:

$$p(\underline{y}|g) = \prod_{k=1}^{d_g} \frac{(r_k)^{r_k\, m_k}}{\Gamma(r_k\, m_k)} \frac{\Gamma(r_k\, m_k + d_k)}{(V_k + r_k)^{r_k\, m_k + d_k}}, \tag{3.5}$$

Furthermore, Bayes theorem gives $p(g|\underline{y}) \propto p(\underline{y}|g)p(g)$, from which $p(g|\underline{y})$ is obtained by normalisation.

Our second aim is to estimate the hazard function, in order to make predictions on survival times. This task can be performed in two steps: first we work conditionally on a partition, and determine a Bayesian estimate of each individual hazard, by calculating the posterior mean $E(\lambda_i|\underline{y}, g)$. Computationally, following (**??**) and (**??**), it turns out that, for $i \in S_k(g)$:

$$E(\lambda_i|\underline{y}, g) = \frac{r_k\, m_k + d_k}{V_k + r_k}, \tag{3.6}$$

The above expression shows that r_k and $r_k\, m_k$ can be interpreted, respectively, as pre-experimental 'total time at risk' and 'observed events' (e.g. coming from a meta-analysis). When no prior information is available, they may be taken in an appropriate *uninformative* manner. The second step of the estimation procedure involves using $p(g|\underline{y})$ to calculate the marginal posterior expectation of each individual hazard $E(\lambda_i|\underline{y})$, via the law of total probabilities:

$$E(\lambda_i|\underline{y}) = \sum_{g=1}^{G} E(\lambda_i|g, \underline{y})p(g|\underline{y}), \tag{3.7}$$

As shown, for instance, in Raftery (1996), using the marginal posterior expectation via the above model averaging procedure leads to predictions better than those based on conditioning on a single partition, such as that associated to the 'best' model.

What illustrated for the exponential hazard will now be generalised to the counting process framework.

3.1.1 A two step Bayesian counting process lifetime value model

We now present our Bayesian version of the counting process model introduced in Section 2.5.

Our proposal can be written as follow:

$$h(t|w, x) = w[h_0(t) + h_1(t|x)], t \leq 0, \tag{3.8}$$

where $h(t|w,x)$ represents a hazard function that has been modified by the inclusion of a frailty. The frailty random variable, w, is assumed to independent of t and x for all clusters with some parametric distribution with unit mean, usually Gamma (Clayton, 1991), where the unknown variance of w, say η, quantifies the amount of heterogeneity among individuals. For more details refer to Section 2.5. To complete the Bayesian specification of the model, prior distributions are needed for the vector parameter λ_0, and the parameter β and σ_θ^2. It seems natural to assume that $\lambda_0 = (\lambda_{00}, ..., \lambda_{0g})^T$ be independent of (β, σ_θ^2) and , η independent of $(\lambda_0, \beta, \sigma_\theta^2)$. Specifically, for λ_0 we assume independent Gamma priors:

$$\lambda_{0k} \sim Ga(\lambda_{0k}|a_{0k}, b_{0k}), k = 1, 2, ..., g_{ij}, \tag{3.9}$$

where $\frac{a_{0k}}{b_{0k}}$ is the prior expectation for λ_{0k} and $\frac{a_{0k}}{b_{0k}^2}$ is the prior variance, with prior independence assumed across the K_{tk} interval, hence:

$$f(\lambda_{00}, ..., \lambda_{0gij}) = \prod_{k=1}^{g_{ij}} Ga(\lambda_{0k}|a_{0k}, b_{0k}). \tag{3.10}$$

For (β, σ_θ^2) we assume the usual Normal-Inverse Gamma conjugate priors, i.e.:

$$\beta|\sigma_\theta^2 \sim N_p(\beta|m_\theta, \sigma_\theta^2 V_\theta), \tag{3.11}$$

with

$$\sigma_\theta^2 \sim Ga(\frac{1}{\sigma_\theta^2}|a_\theta, b_\theta). \tag{3.12}$$

Finally we suggest a Gamma distribution as a prior for η, i.e.:

$$\eta \sim Ga(\phi_1, \phi_2), \tag{3.13}$$

where $\frac{\phi_1}{\phi_2}$ is the prior expectation for η, and $\frac{\phi_1}{\phi_2^2}$ is the prior variance. As discussed in Section 5.2.1 we take $E(\eta) = \frac{\phi_1}{\phi_2} = 1$.

To obtain the conditional posterior distributions, required for Gibbs sampling we use the approach of "data augmentation" (Tanner and Wong, 1987). The idea of data augmentation is to insert latent data or missing data, in order to exploit the simplicity of the resulting conditional posterior distributions of vector parameters of interest. Although this will be increase the dimensionality of the problem (possibly at the expense of extra computing time), the Gibbs sampler will be simpler.

Our objective is to derive the posterior distribution of $(\lambda_0, \beta, \sigma_\theta^2, w)$. Such a posterior can not be computed analytically and, therefore, we used the Monte Carlo

approximations through the Gibbs sampler algorithm. We remark that under the specifications, in Section 2.5, the likelihood it is essentially Poisson in form, reflecting the fact that the likelihood may be thought of as generated by independent contributions of many data each concerned with observation of individual i of cluster j over a very short interval during which the intensity may be regarded as constant, i.e.,

$$P(N_{ijk} = n|dH_{0k}, x_{ij}, \theta_{ij}, w_i, Y_{ij} = 1) = \frac{\left\{w_i \left[dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1})\right]\right\}^n}{n!}$$
$$\times exp\left(-w_i \left[dH0k + \frac{1}{\theta_{ij}}(a_k - a_{k-1})\right]\right). \tag{3.14}$$

In a compact form we have:

$$dN_{ijk} \sim Poisson[dN_{ijk}|w_i dH_{0k} + \frac{w_i}{\theta_{ij}}(a_k - a_{k-1})]. \tag{3.15}$$

Since the additive form of the Poisson sum does not result in the conditional posterior distribution in a closed form, we can solve this problem by expressing the likelihood in an augmented form, involving independent Poisson latent variables, unobserved or missing data, corresponding to each term in the expression for the Poisson mean. In particular, we assume $dN_{ijk} = dN_{ijk0} + dN_{ijk2} + dN_{ijk1}$, for all $i, j : Y_{ij} = 1$ and $k = 1, ..., g$ such that:

$$dN_{ijk0} \sim Poisson[dN_{ijk0}|w_i \lambda_{00}(a_k - a_{k-1})],$$
$$dN_{ijk2} \sim Poisson[dN_{ijk2}|\frac{w_i}{2}\lambda_{0k}(a_k - a_{k-1})^2]$$

and

$$dN_{ijk1} \sim Poisson[dN_{ijk1}|\frac{w_i}{\theta_{ij}}(a_k - a_{k-1})]. \tag{3.16}$$

Using the property that the sum of independent Poisson random variables is also Poisson, it is straightforward to show that the previous equations are equivalent. Such expression allows us to take advantage of Poisson-Gamma conjugate to obtain the conditional posteriors. Indeed in this work we have obtained the required conditionals. The calculations are presented below.

The joint posterior density of the parameters $(\lambda_0, \beta, \sigma_\theta^2, w)$ and latent variables $(dN_{ijk0}, dN_{ijk2}, dN_{ijk1})$ is proportional to:

$$A_1 \times Ga(\eta|\phi_1, \phi_2)Ga(\lambda_{00}|a_{00}, b_{00})N(\beta|m_\theta, \sigma_\theta^2, V_\theta)Ga(\frac{1}{\sigma_\theta^2}|a_\theta, b_\theta), \tag{3.17}$$

where A_1 is:

$$\begin{aligned}
A_1 &= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\prod_{k=1}^{g_{ij}} I(dN_{ijk} = dN_{ijk0} + dN_{ijk2} + dN_{ijk1}) \\
&\times Poisson[dN_{ijk0}|w_i\lambda_{00}(a_k - a_{k-1})] \\
&\times Poisson[dN_{ijk2}|\frac{w_i}{2}\lambda_{0k}(a_k - a_{k-1})^2] \\
&\times Poisson[dN_{ijk1}|\frac{w_i}{\theta_{ij}}(a_k - a_{k-1})] \\
&\times N(log\theta_{ij}|\beta x_{ij}, \sigma_\theta^2)Ga(\lambda_{0k}|a_{0k}, b_{0k})Ga(w_i|\eta^{-1}, \eta^{-1}) \qquad (3.18)
\end{aligned}$$

The sampler iterates through the following steps:

- Step 1. Sample the latent variables $(dN_{ijk0}, dN_{ijk2}, dN_{ijk1})^T$, for all i, j, k : $Y_{ij} = 1$, jointly from their full conditional posterior distribution as follows:

  1. If $dN_{ijk} = 0$, then let $dN_{ijk2} = dN_{ijk0} = dN_{ijk1} = 0$;

  2. If $dN_{ijk} > 0$, then sample $(dN_{ijk0}, dN_{ijk2}, dN_{ijk1})$ from a distribution Multinomial $(dN_{ijk}|P_{ijk0}, P_{ijk2}, P_{ijk1})$ where:

$$\begin{aligned}
P_{ijk0} &= \frac{\lambda_{00}(a_k - a_{k-1})}{\lambda_{00}(a_k - a_{k-1}) + \frac{1}{2}\lambda_{0k}(a_k - a_{k-1})^2 + \frac{w_i}{\theta_{ij}}(a_k - a_{k-1})}, \\
P_{ijk2} &= \frac{\frac{1}{2}\lambda_{0k}(a_k - a_{k-1})^2}{\lambda_{00}(a_k - a_{k-1}) + \frac{1}{2}\lambda_{0k}(a_k - a_{k-1})^2 + \frac{w_i}{\theta_{ij}}(a_k - a_{k-1})}, \\
P_{ijk1} &= \frac{\frac{(a_k - a_{k-1})}{\theta_{ij}}}{\lambda_{00}(a_k - a_{k-1}) + \frac{1}{2}\lambda_{0k}(a_k - a_{k-1})^2 + \frac{w_i}{\theta_{ij}}(a_k - a_{k-1})}. \qquad (3.19)
\end{aligned}$$

  It follows from A_1 that the full conditional distribution of the latent variables is proportional to:

$$\begin{aligned}
A_2 &= I(dN_{ijk} = dN_{ijk0} + dN_{ijk2} + dN_{ijk1}) \times \frac{[w_i\lambda_{00}(a_k - a_{k-1})]^{dN_{ijk0}}}{dN_{ijk0}!} \\
&\times \frac{[(1/2)w_i(a_k - a_{k-1})^2]^{dN_{ijk2}}}{dN_{ijk2}!} \times \frac{[(w_i/\theta_{ij}) \times (a_k - a_{k-1})]^{dN_{ijk1}}}{dN_{ijk1}!}. \qquad (3.20)
\end{aligned}$$

  On the other hand, given A_1 we have A_2 is also proportional to:

$$\begin{aligned}
&\frac{dN_{ijk0}!}{[w_i\lambda_{00}(a_k - a_{k-1}) + [(1/2)w_i(a_k - a_{k-1})^2\lambda_{0k} + (w_i/\theta_{ij})(a_k - a_{k-1})]} \\
&\times \frac{[w_i\lambda_{00}(a_k - a_{k-1})]^{dN_{ijk0}}}{dN_{ijk0}!} \times \frac{[(1/2)w_i\lambda_{0k}(a_k - a_{k-1})^2]^{dN_{ijk2}}}{dN_{ijk2}!} \\
&\times \frac{[(w_i/\theta_{ij}) \times (a_k - a_{k-1})]^{dN_{ijk1}}}{dN_{ijk1}!}, \\
&\propto \frac{dN_{ijk}!}{dN_{ijk0}!dN_{ijk2}!dN_{ijk1}!}P_{ijk0}^d N_{ijk0}P_{ijk2}^d N_{ijk2}P_{ijk1}^d N_{ijk1}, \\
, \quad &\propto Multinomial(dN_{ijk0}, dN_{ijk2}, dN_{ijk1}|dN_{ijk}, P_{ijk0}, P_{ijk2}, P_{ijk1}),
\end{aligned}$$

where $P_{ijk0}, P_{ijk2}, P_{ijk1}$ are defined before.

- Step 2. Sample λ_{00} the full conditional distribution that is proportional to:

$$\begin{aligned}
&\left\{ \prod_{i,j,k:Y_{ij}=1} \frac{[w_i\lambda_{00}(a_k - a_{k-1})]^{dN_{ijk0}}}{dN_{ijk0}!} exp[-w_i\lambda_{00}(a_k - a_{k-1})] \right\} \\
&\times(\lambda_{00})^{a_{00}-1} exp(-\lambda_{00}b_{00}), \\
&\propto (\lambda_{00})^{\sum_{i,j,k:Y_{ij}=1} dN_{ijk0}} exp\left[-\lambda_{00} \sum_{i,j,k} Y_{ij} w_i (a_k - a_{k-1}) \right] \\
&\propto Ga\left\{ \lambda_{00} | a_{00} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{g_{ij}} dN_{ijk0}, b_{00} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{g_{ij}} Y_{ij} w_i (a_k - a_{k-1}) \right\}.
\end{aligned}$$

- Step 3. Sample λ_{0k}, $K = 1, ..., g_{ij}$ and the full conditional distribution of λ_{0k} is proportional to:

$$\begin{aligned}
&\prod_{i,j,k:Y_{ij}=1} [(1/2)w_i(a_k - a_{k-1})^2\lambda_{0k}]^d N_{ijk} exp[-(w_i/2)\lambda_{0k}(a_k - a_{k-1})^2] \lambda_{0k}^{a_{0k}-1} \\
&\times exp(-\lambda_{0k}b_{0k}), \\
&\propto (\lambda_{0k})^{\sum_{i=1}^{n}\sum_{j=1}^{m_i} dN_{ijk}} exp\left[-(w_i/2)\lambda_{0k} \sum_{i=1}^{n}\sum_{j=1}^{m_i} Y_{ij}(a_k - a_{k-1}) \right] \\
&\propto \lambda_{0k})^{a_{0k}-1} exp(-\lambda_{0k}b_{0k}) \\
&\propto Ga\left(\lambda_{0k} | a_{0k} + \sum_{i,j} d_{ijk}, b_{0k} + (\frac{w_i}{2}) \sum_{ij} Y_{ij}(a_k - a_{k-1}) \right).
\end{aligned}$$

- Step 4. To derive the conditional distribution of w_i , $i = 1, ..., n$ we start with the joint posterior density of parameters prior to augmentation that is proportional to:

$$\begin{aligned}
&\prod_{j=1}^{m_i} \prod_{Y_{ij}=1} \left\{ w_i \left[dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right] \right\}^{dN_{ijk}} exp\left\{ -w_i \left[dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right] \right\} \\
&\times w_i^{\eta^{-1}} exp\left(-\eta^{-1} w_i \right), \\
&\propto (w_i)^{\sum_{j=1}^{m_i}\sum_{k=1}^{g_{ij}} dN_{ijk} + \eta^{-1} - 1} exp\left\{ -w_i \left(\eta^{-1} + \sum_{j=1}^{m_i}\sum_{k=1}^{g_{ij}} dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right) \right\}, \\
&\propto Ga\left(\sum_{j=1}^{m_i}\sum_{k=1}^{g_{ij}} d_{ijk} + \eta^{-1}, \sum_{j=1}^{m_i}\sum_{k=1}^{g_{ij}} dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right).
\end{aligned}$$

- Step 5. The full conditional distribution of (β, σ_θ^2) is proportional to:

$$\left\{ \prod_{i=1}^{n}\prod_{j=1}^{m_i} N(log\theta_{ij} | \beta x_{ij}, \sigma_\theta^2) \right\} N(\beta | m_\theta, \sigma_\theta^2, V_\theta) Ga\left(\frac{1}{\sigma_\theta^2} | a_\theta, b_\theta \right). \qquad (3.21)$$

That expression is the same as one that appear in the usual conjugate analysis of the Normal data (see e.g. DeGroot, 1970). It is then proportional to a Multivariate Normal Inverse Gamma distribution, i.e.,

$$\begin{aligned} &\beta|\sigma_\theta^2 \sim N_p(\beta|\hat{\beta}, \sigma_\theta^2(V_\theta^{-1} + xx^T)), \\ &\sigma_\theta^2 \sim Ga\left(\frac{1}{\sigma_\theta^2}|a_\theta + \frac{V}{2}, b_\theta + \frac{1}{2}[(y - x\hat{\beta})^T y + (m_\theta - \hat{\beta})^T V_\theta^{-1} m_\theta\right), \end{aligned} \tag{3.22}$$

where $y = (log(\theta_{11}), ..., log(\theta_{nm})^T$, x is the covariate matrix and the estimated of the coefficient of regression β is calculated form $\hat{\beta} = (V_\theta^{-1} + x^T x)^{-1}(V_\theta^{-1} m_\theta + x^T y)$. The other conditionals do not have a conjugate analysis. For each, $j = 1, ..., m_i$ and $i = 1, ..., n$ the conditional distribution for θij is proportional to:

$$Y_{ij}(t) w_i [h(t|\lambda_0, \theta_{ij})]^{dN_{ijk}} S(t|\lambda_0, \theta_{ij}) F(\theta_{ij}|\beta x_{ij}, \sigma_\theta^2), K = 1, ..., g_{ij}. \tag{3.23}$$

This expression does not have closed form. But it is still possible to sample from it using a Metropolis algorithm.

Finally, for $i = 1, ..., n$, letting $\xi = \eta^{-1}$, the full conditional distribution of ξ does not have a closed form either. It is proportional to:

$$\left(\prod_{i=1}^{n} w_i^{\xi - 1}\right) \xi^{-n\xi} \frac{exp\left(-\xi \sum_{i=1}^{n} w_i\right)}{[\Gamma(\xi)]^n} f(\xi), \tag{3.24}$$

with $\xi \sim Ga(\xi|\phi_1, \phi_2)$. With this choice of prior it can be shown that the above full conditional density is log-concave. Thus we can use the adaptive rejection algorithm to sample from this full conditional. The results are in the application Section.

3.2 Bayesian Model Averaging for variable and model selection

As we have discussed in Chapter 1, variable selection has been recognized as "one of the most pervasive model selection problems in statistical applications" (George et al., 2000), and a lot of different methods emerged during the last 30 years, especially in the context of linear regression (see Miller 1990, McQuarrie and Tsai, 1998, George et al., 2000). Many researchers focused on developing an appropriate model selection criterion assuming that few reasonable models are available, such as, Mallows Cp, (Mallows, 1973), BIC (Schwarz 1978), RIC (Foster and George 1994), bootstrap model selection (Shao, 1996). However in reality the researchers often

have to choose a single or few best models from the enormous amount of potential models using techniques such as the stepwise regression of Efroymson (1960) and its different variations, or, for example, the leaps-and-bounds algorithm of Furnival and Wilson (1974). Typically researchers use both approaches, first trying to generate several best models for different numbers of variables, and then select the best dimensionality according to one of the criteria listed. Any combination of these approaches to model selection, however, do not seem to take into account the uncertainty associated with model selection and therefore in practice tend to produce overoptimistic and biased prediction intervals, as will be discussed later. In addition, the statistical validity of various variable selection and elimination techniques (stepwise and forward selection, backward elimination) is suspect. The computations are typically organized in "one variable at a time" fashion seemingly employing the statistical theory of comparing two nested hypotheses, however ignoring the fact that the true null distributions of the widely used "F statistics" (such as F-to-enter) are unknown and can be far from the assumed F distribution (see Miller, 1990).

The two sides of the model selection problem (model search and model selection criterion) are naturally integrated in model averaging which overcomes the inherent deficiency of the deterministic model selection by combining (averaging) information on all or a subset of models when making estimation, inference, or prediction, instead of using only one model. In this review we will focus on a standard Bayesian approach to model selection which associates a prior probability with each model M in some model space M (see Key et al., 1999 for differences between the $M-close$, $M-open$ and $M-completed$ perspectives to modelling) and then uses their posterior probabilities to select one best or "several best" models (for a discussion of different approaches to Bayesian model selection see e.g. Kass and Raftery, 1995). Bayesian Model Averaging (BMA) goes further and uses these probabilities to average the "model parameter" when computing the posterior probabilities associated with the other parameters, nested within the model. BMA is becoming an increasingly popular data analysis tool which allows the data analyst to account for uncertainty associated with the model selection process. In this review, we will try to think about models in a broad context when appropriate since different researchers applied BMA within quite different classes of models.

As is clear from the above discussion, BMA arises when the true model is unknown

before we look at the data (actually it assumes that there may be no single true model) and it can be viewed as a data analysis tool which in a sense brings together the exploratory phase of the data analysis (model specification) and the confirmatory phase (model estimation) by simultaneously searching through data for good models and updating their associated posterior probabilities (if necessary). Then BMA combines the predictions and parameter estimates obtained with different plausible models using their posterior probabilities as weights. A popular part of the BMA output is variable assessment which can be done by aggregating posterior weights across only those models where a given variable was present. More technically, following Madigan and Raftery (1994), if Δ is the quantity of interest, such as a parameter of the regression model or a future observation, then its posterior distribution given data D and a set of K models is a mixture of posterior distributions (see also Leamer, 1978):

$$p(\Delta|D) = \sum_{k=1}^{K} pr(\Delta|M_k, D)pr(M_k|D), \tag{3.25}$$

the posterior probability for model M_k is given by:

$$p(M_k|D) = \frac{pr(D|M_k)pr(M_k)}{\sum_{l=1}^{K} pr(D|Ml)pr(M_l)}, \tag{3.26}$$

where

$$pr(D|M_k) = \int pr(D|\theta_k, M_k)pr(\theta_k|M_k)d\theta_k, \tag{3.27}$$

is the predictive distribution for model M_k.

The issues that arise when implementing BMA can summarized as follows:

- assigning prior distributions for different models and model parameters
- searching through the model space for data-supported models
- computing posterior model probabilities
- drawing inference in BMA: obtaining estimates and probability/confidence intervals for the parameters and observables
- measuring predictive performance of BMA

The topic of selecting the best model has received a considerable attention and generated enormous literature in statistics (see for example Miller, 1990, where selecting

subsets in regression is considered). Now it seems that the challenging problem of finding the best subset of variables has somewhat obscured and overshadowed a not less important issue of aggregating many good models even in the idealized situation when the best subsets (with respect to some criterion) can be trivially found. It is interesting to note that many researchers ignored model uncertainty even when the information on model selection uncertainty was a natural by product of the proposed methods, such as for example, distribution of different models in multiple runs of stepwise regression with random starting subsets (see Miller, 1990), or the output of the leaps and bounds algorithm proposed in Furnival and Wilson (1974). Searching for good models is also an integral part of BMA, which "simply" averages across the complete model space. Therefore different approaches were proposed for implementing the BMA methodology on a reduced model space. The idea is to search through model space for most reasonable models, that is models supported by the data (Madigan and Raftery, 1994). Two different approaches were proposed in literature, deterministic search and stochastic search on the model space using Markov chain Monte Carlo (MCMC).

3.2.1 Deterministic model search

The suggested BMA deterministic search schemes are Occam Window method of Madigan and Raftery (1994) and leaps and bound technique adopted in Volinsky et al. (1997). Occam Window method is described in Hoeting et al. (1999). The idea is to avoid averaging over the complete space of possible models by restricting it to only the models well supported by the data. For instance, consider the class:

$$A' = \left\{ M_i : \frac{max_j pr(M_j|D)}{pr(M_i|D)} \leq C \right\}, \tag{3.28}$$

for some appropriate C, say C=20 (as suggested in Raftery, 1995). This set can be further reduced by eliminating the set B of models that have less posterior support than their own sub-models (Occam razor):

$$B = \left\{ M_i : \exists M_l \in A', M_l \subset M_i, \frac{pr(M_l|D)}{pr(M_i|D)} > 1 \right\}, \tag{3.29}$$

and then, the model averaging is performed over the set $A = \acute{A}|B$. The approximation provided by this method to full BMA was shown to be good in several applications. Criticism for this approach was expressed by several writers (Clyde,

1999a George and Clyde, 1999) that averaging across a relatively small set of models captured by the Occam window may result in biased estimates and predictions. The authors of this approach argue that while it certainly overestimates the posterior probabilities of the models it includes, it seems to preserve the ratios of these probabilities (Hoeting et al., 1999). Also according to them, there is some philosophical justification for the proposed method as it corresponds to the practice of model rejection well established in the scientific community: "models that have been clearly discredited do get discarded in scientific research". As was argued in Madigan and Raftery (1994), averaging across all models assumed in standard Bayesian approach is incorrect, "adopting standard methods of scientific investigation, we contest that accounting for the true model uncertainty involves averaging over a much smaller set of models". Another deterministic way to search for models in A was suggested in Volinsky et al. (1997) and it employs the "leaps and bounds" algorithm generalized for the non-linear case. It may produce results similar to the Occam razor approach, though it apparently lacks the philosophical justification of the latter.

3.2.2 Stochastic model search via MCMC

The first applications of the stochastic model search were seen in the Markov Chain Monte Carlo Model Composition, in Madigan and York (1995), Clyde et al. (1996), and Stochastic Search Variable Selection via Gibbs sampler of George and McCulloch (1993). The Reversible Jump MCMC algorithm of Green (1995) includes many of these algorithms as special cases (see also Carlin and Chib, 1995). This universal approach of MCMC became an extremely popular tool in Bayesian computations. Two aspects of MCMC in the context of BMA are that it is:

- an excellent search device that allows one to locate good models
- a mechanism of computing the posterior probabilities as proportion of visited models

There are two different implementations of MCMC in BMA: when the Markov chain is constructed only in the model space and when it is constructed in the combined model and parameters space.

3.2.3 Stochastic model search in the model space

The MC^3 is a special case of the Metropolis-Hastings (MH) algorithm (Hastings, 1970). The main idea of MH algorithm is to set-up an irreducible and aperiodic Markov chain whose equilibrium distribution is the desired posterior distribution (see Smith and Roberts, 1993). In the case of model selection, this can be done as follows (Hoeting et al., 1999). First the Markov chain is constructed, defining a neighbourhood nbd(M) for each model. Define a transition matrix q by setting $q(M \rightarrow M') = 0$ for $M' \notin nbd(M)$ and $q(M \rightarrow M') \neq 0$, $\forall M' \in nbd(M)$. If the chain is currently in state M, proceed by drawing M' from transition matrix q. M' is accepted with probability:

$$\alpha = min\left\{1, \frac{pr(M'|D)}{pr(M|D)}\right\}, \tag{3.30}$$

Otherwise the chain remains in state M. Under these conditions, the limiting ratio of the number of times each state (model) is visited to the total number of draws is proportional to the posterior model probability. Note that to be able to compute the ratio, we must know the posterior probabilities $pr(M|D)$ only up to normalizing constant, which can be obtained from the Bayes Factor (BF) as follows. Note that the previous expression can be written as:

$$\alpha = min\left\{1, \frac{pr(D|M')pr(M')}{pr(D|M)pr(M)}\right\}, \tag{3.31}$$

where the ratio $p(D|M')/p(D|M)$ is the Bayes Factor for the two models used as a measure of predictive power for a model M' against M. One can see that the outlined procedure is actually a Metropolis algorithm, a special case of the Metropolis-Hasting algorithm, which arises when the proposal probability is symmetric and $q(M \rightarrow M')$ is same as its reverse jump probability, $q(M' \rightarrow M)$. In general, for the M-H algorithm,

$$\alpha = min\left\{1, \frac{pr(M'|D)q(M \rightarrow M')}{pr(M|D)q(M' \rightarrow M)}\right\}, \tag{3.32}$$

The information on model posterior probabilities is accumulated while performing the stochastic search. In many cases, however, MC^3 can be used just as a searching device, since the posterior probabilities can be obtained via the BIC approximation for Bayes Factors (see Noble, 2000). Furthermore, the availability of an analytical approximation for the posterior probabilities makes the convergence properties of the underlying Markov chain less relevant (Clyde et al. 1996; Noble, 2000) and

allows for more efficient mixing by restarting the simulation at random models and carrying out multiple chains of the MCMC (Noble, 2000; Chipman et al., 1998). Noble (2000) combined the Metropolis algorithm with the Occam razor criterion of model screening, which allowed him to reduce the number of candidate models for the final averaging. As was noted in Clyde (1999a), different versions of this algorithm arise with other choices of proposal probability $q(M \rightarrow M')$, which may lead to procedures that can move more rapidly through the model space. Clyde (1999) uses approximate posterior probabilities of variable inclusion for the proposal distribution to target more important variables, rather than proposing all variables to be added or deleted with equal probability $1/p$.

3.3 Stochastic search in the combined model and parameter spaces

Another MCMC approach was proposed in George and McCulloch (1993). The subset search was implemented via the Gibbs sampler and the movement occurred in the combined model and parameter space of the associated Markov chain. Following George and McCulloch (1993), the Gibbs sampler is used to generate a sequence $\gamma^1, ...\gamma^m$, where $\gamma = (\gamma_1, ..., \gamma_p)$ is a vector of ones and zeros corresponding to the presence/absence of a given variable in the model. This sequence converges rapidly in distribution to $pr(\gamma|D)$ (that is $pr(M|D)$ in our more general notation) and with high probability contains most interesting subsets. Those γ with highest probability will appear most frequently and hence will be easier to identify. Furthermore, for the non-conjugate mixture prior, the sequence of models is embedded into the auxiliary Gibbs sequence, which is an ergodic Markov chain:

$$\beta^0, \sigma^0, \gamma^0, \beta^1, \sigma^1, \gamma^1, ..., \beta^j, \sigma^j, \gamma^j, ..., \tag{3.33}$$

where β^0, σ^0 are initialized to be the least squares estimates of model $Y|\beta, \sigma^2$ according to a $N(X\beta, \sigma^2 I)$ and $\gamma^0 = (1, 1, ..., 1)$, while the subsequent values β^j,σ^j,γ^j are obtained by simulating values according to the following iterated sampling scheme:

$$\begin{aligned} \beta^j &\sim pr(\beta^j|Y, \sigma^{j-1}, \gamma^{j-1}), \\ \sigma^j &\sim pr(\sigma^j|D, \beta^j, \gamma^{j-1}), \\ \gamma_i^j &\sim pr(\gamma_i^j|D, \beta^j, \sigma^j), \end{aligned} \tag{3.34}$$

$$\gamma_{(i)}^{j} = pr(\gamma_i^j | \beta^j, \sigma^j, \gamma_{(i)}^j)$$
$$\gamma_{(i)}^{j} = (\gamma_1^j, ..., \gamma_{i-1}^j, \gamma_{i+1}^{j-1}, ..., \gamma_p^{j-1})$$

The authors have to use a continuous prior instead of mass point at $\beta_i = 0$, because otherwise their Gibbs sampler would get stuck each time it generates $\beta_i = 0$. To obviate this problem Geweke (1996) proposed an alternative Gibbs sampler implementation, which jointly draws (β_i, γ_i)one at a time given σ, and the other pairs (β_l, γ_l), $l \neq i$. The Gibbs updating scheme is much simpler and involves only the single sequence $\gamma^1, ..., \gamma^m$, when each γ value can be generated component wise from the full conditionals $\gamma_i|\gamma_{(i)}$, D for $i = 1, ..., p$, where γ_i can be drawn in any fixed or random order. The generation of components in this sequence can be obtained simply as simulations of Bernoulli draws with probabilities that can be easily computed (see details in George and McCulloch, 1997) They also note an interesting fact that for the conjugate prior, their Gibbs sampler is equivalent to a Metropolis algorithm modified as follows. Components of γ are randomly permuted and considered to be added or deleted from the model in turn, rather than selected with equal probability $1/p$, with the acceptance probability given by the ratio,

$$\alpha = \frac{pr(M'|D)}{pr(M|D) + pr(M'|D)}. \tag{3.35}$$

Since α is always < 1, it makes the jump less probable and hence the algorithm is less efficient than the MC^3 however this can be somewhat compensated by the fact that in the Gibbs sampler the components are selected cyclically, which ensures better mixing. A hybrid algorithm was considered in Clyde et al. (1996).

Alternatively, Clyde et al. (1996) derived approximate formulas assuming conditional independence of γ_i, and showed that under orthogonality of predictors (which can be always achieved by an appropriate transformation) MCMC sampling can be replaced by direct importance sampling of the model components element wise. Clyde et al. (1998) further developed a sampling-without-replacement algorithm which can be yet another efficient alternative to MCMC under orthogonality of predictors. Clyde (1999) also considered the block Gibbs sampler with both γ and σ (when σ is not integrated out) and showed how to implement a very efficient blocked Gibbs sampler, where $\gamma|D$, σ^2 is distributed exactly as a product of independent Bernoulli distributions and $\sigma^2|\gamma, D$ has the inverse gamma distribution (again assuming orthogonality of predictors). Instead of integrating σ out, the Rao-

Blackwellized estimator of marginal $pr(\gamma_i = \frac{1}{D})$ and posterior expectation of model parameters β can be obtained by averaging it over the values of σ^2 from the Gibbs sampler. Clyde (1999) extended this result for generalized linear models by applying the appropriate variance-stabilizing transformation and using the approximation for the posterior model probabilities as a product of Bernoulli distributions. When the predictors are correlated, Clyde et al. (1996) proposed using orthogonalization of the original model space similar to principal component transformation.

3.3.1 Computing Model Posterior probabilities

To compute the posterior probabilities we have to evaluate the marginal likelihood $pr(D|M_k)$. For certain models such as linear regression with normal errors and conjugate priors, the expression for the marginal likelihood of model M_k is readily available (see Raftery, 1996; Hoeting et al., 1999). In other cases one can use analytical approximations.

Kass and Wasserman (1995), Kass and Raftery (1995), Raftery (1996) demonstrated various approximations that can be obtained via the Laplace method. The Bayesian Information Criterion (BIC, Schwarz, 1978) approximation is the simplest.

$$logpr(D|M_k) \sim logpr(D|\hat{\theta}_k, M_k) - (p_k/2)log(n), \tag{3.36}$$

where p_k is the number of parameters in the model M_k.This approximation gives us the expression for the Bayes Factor, $B_{10} = \frac{pr(D|M_1)}{pr(D|M_0)}$ as:

$$2logB_{10} \sim \chi^2 - (p_1 - p_0)log(n), \tag{3.37}$$

where $\chi^2 = 2\left\{logpr(D|\hat{\theta}_1, M_1) - logpr(D|\hat{\theta}_0, M_0)\right\}$, the standard likelihood ratio test statistic when M_0 is nested within M_1, p_0 and p_1, are the number of parameters in M_0 and M_1.

According to the authors, although this approximation is least accurate and should be of order $O(1)$, however practical experience suggests that it performs surprisingly well, as if it were of order $O(n^{-1/2})$, see Kass and Wasserman (1995) for justifications. Note that the BIC approximation is indeed of order $O(n^{-1/2})$ for a certain prior, namely the Unit Information Prior, described in the previous Section (see derivation in Raftery, 1995) and therefore using BIC though it does not require to specify the priors for the model parameters, implicitly assumes UIP. Now, if we want to obtain

the posterior probabilities by averaging with respect to a selected set of $(k+1)$ not necessarily nested models, say, we can compare models $M_1 ... M_k$ to M_0 in turn and compute:

$$pr(M_k|D) = \frac{\alpha_k B_{k0}}{\sum_{r=0}^{K} \alpha_r B_{r0}}, \tag{3.38}$$

where $\alpha_k = \frac{pr(M_k)}{pr(M_0)}$ is the prior odds for M_k against M_0 (renormalization of unnormalized posterior probabilities, Clyde et al., 1996).

Estimating posterior probabilities $pr(M_k|D)$ via MC^3 or related methods of course would allow us to estimate these probabilities empirically as the proportion of times it was visited during the MC^3 simulation, however for computing the acceptance rule in the MC^3 algorithm we still need ratios of posterior probabilities for models M_1 and M_0, say, which can be trivially obtained as $B_{10}\alpha_1$. As was observed in Noble (2000), different forms of the model choice criteria can be translated into the standard BIC approximation by the corresponding adjustment of the prior model probabilities, so that the difference in the criteria is absorbed by the α's. This can be considered yet another way of calibrating the model priors.

3.3.2 Using MCMC approaches to compute Bayes Factors

What shall we do if the Laplace approximation to the Bayes Factors does not work? Then we cannot use MC^3, which requires the ratio of the posterior model probabilities, however we can resort to other MCMC techniques where the Markov chain is defined on the combined space of model and model parameters. The apparent problem is that moving across models with different composition of parameters may render the associated Markov chain not irreducible in the general case.

George and McCulloch (1993) handled it by artificially assigning non-zero priors to the parameters that were supposed to be deleted from the model. Carlin and Chib (1995) proposed a general but apparently quite wasteful Gibbs sampling scheme by introducing so-called pseudo-priors (or linking densities) that were linking the parameters missing in a given model with that model. The most general approach that allows one to implement the MCMC for the models with variable parameter spaces was introduced in Green (1995) who proposed his Reversible Markov chain Monte Carlo approach. In general its implementation requires evaluation of the Jacobians associated with transformations from one parameter space to another. The Monte

Carlo estimate for the model probabilities is based on their visiting frequency. Note, however, that when the number of parameters is too large, the MCMC algorithm may present convergence problems and may require special efforts for the problem specific tuning. Lewis and Raftery (1997) proposed a combination of MCMC and Laplace approximation for the Bayes factor which they called the Laplace-Metropolis estimator. This method uses the output of Gibbs sampler to estimate the parameters of the M_1 and M_0 at their posterior modes and then plugs it into the expressions for the likelihood and the inverse Hessian necessary for the Laplace method.
Clyde (1999) found that in the case of the normal linear model with orthogonal predictors and known σ (orthogonalization can be accomplished by principal components transformation, also regression with orthogonal predictors becomes of central importance in nonparametric wavelet estimation, see Clyde and George, 1999), it is possible to obtain the exact analytical expression for the posterior odds when the relationship between Y and the predictors is linear. Clyde (1999) also explored an alternative way to approximate the model probabilities via the log-linear representation of the model space.

3.3.3 Inference with Fully Bayesian Model averaging

We can distinguish between "fully Bayesian" analysis when both parameters and models are treated in a Bayesian way and semi-Bayesian approach when the estimates of the model parameters are obtained using the classical methods and only the model space is treated in a Bayesian way (Volinsky et al., 1997; Noble, 2000). In this case, we either obtain both posterior probabilities and parameter estimates from MCMC (George and McCulloch, 1993), or use the explicit formulae for the posterior distributions of the model parameters, if available (see Raftery et al., 1997, for example). In either case, using MCMC is a standard way of obtaining approximate model posterior probabilities. For the approach the "model averaged" estimate of a quantity of interest can be obtained in general as follows. Let $g(M_i)$ be defined on model space M, then the average:

$$\hat{G} = \frac{1}{N}\sum_{t=1}^{N} g(M(t)), \tag{3.39}$$

is an estimate of $E[g(M)]$. $\hat{G} \to E(g(M))$ a.s. as $N \to \infty$, and we can compute a posterior distribution by letting $g(M) = pr(\Delta|M, D)$ (Madigan and Raftery, 1994).

George (1999a) pointed out that when $pr(D|M_k)$ can be either computed exactly or approximated very well (using the BIC approximation, for example), we can compute $pr(M_k|D)$ and approximate as:

$$pr(\Delta|D) = \sum_{M_k \in S} pr(\Delta|M_k, D) pr(M_k|D), \tag{3.40}$$

over the selected subset of models S rather than using the Monte Carlo estimate. In the same fashion we can compute, say

$$E(\Delta|D) = \sum_{M_k \in S} E(\Delta|M_k, D) pr(M_k|D) \tag{3.41}$$

or expectation of any other quantity of interest. Note that sometimes we can use the ML approximation for the posterior probability, essentially providing a link from the fully Bayesian to semi-Bayesian model averaging (Volinsky et al., 1997).

$$pr(\Delta|M_k, D) \sim pr(\Delta|M_k, \hat{\theta_k}, D) \tag{3.42}$$

In the semi-Bayesian approach, it is only the model space that is treated in a Bayesian way. We can still use MCMC or Occam window to search for models and use BIC approximation to the Bayes Factors (or some of its generalizations, see Smith and Spiegelhalter, 1980; Noble, 2000) to obtain model weights over the set of the selected models. (Note that the BIC does not require specifying the priors for the model parameters.) Using the model weights, we can aggregate parameter estimates and predictions obtained by applying classical methods to fitting these models. Averaging parameter estimates across the model space can be interpreted as shrinking estimates toward zero when the variable is not active in the majority of models. This procedure sometimes generates criticism, as to its validity. It seems that combining, say, regression coefficient for the same variable from different models is inappropriate because they are adjusted for different predictors (Draper, 1995) and hence not directly comparable. Hoeting et al. (1999), although defending the general weighting approach, advises not to abuse it and apply it mostly to observables, such as predicted future observations rather than to the individual coefficients. George and McCulloch (1997) also do not advise to go any further in averaging than combining predictions from different models using posterior model probabilities.

3.4 A one step Bayesian lifetime value model

Methods for analyzing survival data often focus on modelling the hazard rate. The most popular way of doing this is to use the Cox proportional hazards model (Cox, 1972) which allows different hazard rates for cases with different covariate vectors and leaves the underlying common baseline hazard rate for subject i with covariate vector X_i as specified in (5).

Since the integrals required for BMA do not have a closed-form solution for Cox models, Raftery, Madigan and Volinsky (1995) and Volinsky et. al (1997) adopted a number of approximations for Bayesian Model averaging. In particular it is possible to use the MLE approximations,

$$p(\Delta|M_k, D) \approx p(\Delta|M_k, \hat{\beta}_k, D), \tag{3.43}$$

and the Laplace approximation,

$$\log p(D|M_k) \approx \log p(D|\hat{\beta}_k, M_k) - d_k \log(n), \tag{3.44}$$

where d_k is the dimension of β_k and n is usually taken to be the total number of cases. This is the Bayesian Information Criterion (BIC) approximation. Volinsky et al. (1997) provides evidence that n should be the total number of *uncensored* cases (i.e. events), as we discussed in Section 5.3.

To implement BMA for Cox Models, we have followed Raftery et. al (1999) and used an approach similar to the Occam's window method, implemented in a set of R routines allowed for Bayesian Model Averaging. To efficiently identify good models, we adapt the 'leaps and bounds' algorithm of Furnival and Wilson (1974) which was originally created for linear regression model selection. The leaps and bounds algorithm provides the top q models of each model size, where q is designated by the user, plus the MLE $\hat{\beta}_k$, $var(\hat{\beta}_k)$, and R_k^2 for each model M_k are returned. Lawless and Singhal (1978) and Kuk (1984) provided a modified algorithm for no-normal regression models that gives an approximate likelihood ratio test statistic and hence an approximate BIC value. With BMA it is possible to have for each model a BIC, the posterior probability and for each parameter the relative mean, the variance and also the posterior probability that a Cox regression coefficient for a variable is nonzero ('posterior effect probability') as the sum of posterior probabilities of the models which contain that variable.

The posterior mean, follow Raftery et al. (1999) of a regression coefficient can be shown to be:

$$\begin{aligned}\hat{\theta}_{BMA} &= E_M(\hat{\theta}) = \sum_{i=1}^{K} \hat{\theta}_i P(M_i|D) \\ &= \frac{\sum_{i=1}^{K} \hat{\theta}_i P(M_i|D)}{\sum_{i:\theta_i \in M_i} P(M_i|D)} \times \sum_{i:\theta_i \in M_i} P(M_i|D) \\ &= E(\hat{\theta}|\theta_i \in M_i) \times P(\theta \neq 0), \end{aligned} \tag{3.45}$$

which is the conditional posterior mean of θ multiplied by its posterior probability. Similarly, to calculate the variance of the regression coefficient, let $p_i = P(M_i|D)$ and $V_i = Var(\hat{\theta}|M_i, D)$. Then:

$$\begin{aligned} V(\hat{\theta}) &= E(\hat{\theta}^2) - (\sum_{i=1}^{K} p_i \hat{\theta}_i)^2 \\ &= \sum_{i=1}^{K} p_i (V_i + \theta_i^2) - (\sum_{i=1}^{K} p_i \hat{\theta}_i)^2 \\ &= \sum_{i=1}^{K} p_i V_i + [\sum_{i=1}^{K} p_i \hat{\theta}_i^2 - (\sum_{i=1}^{K} p_i \hat{\theta}_i)^2] \\ &= \sum_{i=1}^{K} p_i V_i + \sum_{i=1}^{K} p_i (\hat{\theta}_i - \sum_{i=1}^{K} p_i \hat{\theta}_i)^2, \end{aligned} \tag{3.46}$$

Note that the first term is the weighted variance over models, but the overall variance is affected by the second term, which depends on how stable the estimates are across models. The more these estimates differ across models, the higher the posterior variance. In this way the standard errors reported for variables directly take into account model uncertainty.

Prior probabilities on both model space and parameter space are defined by this procedure. All models are considered likely *a priori* by the leaps and bounds algorithm. Using the BIC approximation to the integrated likelihood defines an implicit prior on all the regression parameters, as described before.

When prior information about the importance of a variable is available for model structures a prior probability for a model M_i can be specified as

$$p(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} (1 - \pi_j)^{1-\delta_{ij}}, \tag{3.47}$$

where $\pi_j \in [0, 1]$ is the prior probability that $\beta_j \neq 0$ in a regression model and δ_{ij} is an indicator of whether or not variable j is included in model M_i. Assigning $\pi_j = 0.5$

for all j corresponds to a uniform prior across model space, while $\pi_j < 0.5$ for all j imposes a penalty for large models. Using $\pi_j = 1$ ensures that variable j is included in all models.

This approach was used to specify model priors for variable selection in linear regression in George and McCulloch (1993) and suggested for model priors for BMA in Cox models by Raftery et al. (1999).

3.4.1 Bayesian lifetime value model adequacy

In order to evaluate model adequacy we now derive the BIC for censored survival models.

When censoring is present it is unclear whether the penalty in BIC should use n, the numbers of observations, or d, the number of events. When using the partial likelihood (Cox, 1972) there are only as many terms in the partial likelihood as there are events d. Kass and Wasserman (1995) indicate that the term used in the penalty should be the rate at which the Hessian matrix of the log-likelihood function grows, which suggest that d is the correct quantity to use. However, if we are to use a revised version of BIC, it is important that the new criterion continue to have the asymptotic properties that Kass and Wasserman derived.

In our proposal we use a revised BIC (Volinsky et al. 1999) with a slightly modified outcome. Let us alter condition 3 to be:

$$-\frac{1}{d}D^2 l(\hat{\theta}\hat{\psi}) - I_u(\theta,\psi) = O_p(n^{-\frac{1}{2}}), \tag{3.48}$$

where $I_u(\theta,\psi)$ is the expected Fisher information for one uncensored observation, (the uncensored unit information). If this holds, then 3.37 is true, and the new BIC (with d in the penalty) is an $O_p(n^{-\frac{1}{2}})$ approximation to twice the Bayes factor where the prior variance on θ is now equal to the inverse of the uncensored unit information. By using d in the penalty instead of n, it can be shown that this asymptotic result holds, the only difference being in the implicit prior on the parameter. Indeed (Kass and Wasserman, 1995) argue that under null orthogonality and independent censoring, equation (3.37) holds for Cox models when d is used in the penalty for BIC, and the Normal overall unit information prior is used.

3.4.2 Bayesian lifetime value variable selection

When BMA is applied to the subset selection problem, its output can be used to rank the variables in order of their importance so as to obtain a single model including the best variables, if needed. This procedure was singled out as Bayesian Variable Assessment (BVA, the term seems to be first used in Meyer and Wilkinson, 1998) and it produces variable activation probabilities (weights), obtained by simply averaging out the posterior model probabilities into a variable weight over the set of models were it is present.

When the predictors are orthogonal, Clyde (1999) and Clyde et al. (1996) showed how to directly obtain the exact or approximate posterior probabilities for variables to be included in the model. These activation probabilities, if known before the model search, can be used in MCMC as proposal probabilities. The posterior effect probabilities, $pr(\beta \neq 0|D)$ can be compared to the classical $P - value$. Also $P - values$ cannot distinguish between the situation when we fail to reject the null hypothesis because of insufficient data and when the data actually provide evidence for H_0 (Hoeting et al., 1999). The authors in Hoeting et al. (1999) speculate that their posterior effect probabilities can be associated with scientific significance, as opposed to statistical significance which can sometimes lead to false conclusions. Finally, we can use variable assessment for building a single model that includes only top variables according to their weights. Such a single model approach would certainly violate the philosophy of BMA, however it can be thought of as one of it possible outputs. A more subjective interpretation of the model parameters and the associated effect probabilities was proposed in Mitchell and Beauchamp (1988), Geweke (1996), Laud and Ibrahim (1995, 1996). In their approach, there is not much physical meaning that can be ascribed to the model parameters, only observable quantities are interpretable. According to Mitchell and Beauchamp (1988), the prior probability, $pr(\beta_i \neq 0)$ represents the proportion of credible experts who would include the variable in the model. The predictive approach to BMA is also inherent in Clyde et al. (1996), since the parameters associated with their orthogonalized predictors clearly do not carry much physical meaning.

3.5 Application of Lifetime Value Models

Our case study concerns a media service company. For non disclosure reasons in this paper we cannot give accurate statements and information about the company whose data we have analysed; we shall instead use general statements and normalised figures; the company will be referred to as 'the company'.
The main objective of such a company is to maintain its customers, in an increasingly competitive market, to evaluate the lifetime value of such customers, to carefully design appropriate marketing actions. The company is such that most of its sales of services are arranged through a yearly contract, which allows buying different 'packages' of services at different costs. The contract of each customer with the company is thus renewed yearly. If the client does not withdraw, the contract is renewed automatically. Otherwise the client churns. In the company there are three types of churn events: people that withdraw from their contract in due time (i.e. less than 60 days before the due date); people that withdraw from their contracts overtime (i.e. more than 60 days before the due date); people that withdraw without giving notice, as is the case of bad payers. Correspondingly, the company assigns two different churn states: an 'EXIT' state to the first two classes of customers; and a 'SUSPENSION' state to the third.
Concerning the causes of churn, it is possible to identify a number of components that can generate such behaviour:

- A static component, determined by the characteristics of the customers and the type/subject of contracts;
- A dynamic component, that encloses trend and the contacts of the clients with the call center of the company;
- A seasonal part, tied to the period of subscription of the contract;
- External factors, that include the course of the markets and of the competitors.

Currently the company uses a classification tree model that gives, for each customer, a probability of churn (score). The goal for the company is to identify customers that are likely to leave and join a competitor.
In business terms, predictive accuracy means being able to identify correctly those

individuals who will become really churner during the valuation phase (correct identification). Evaluation can be made using a confusion, or cross validation matrix. Static models, such as classification trees, show an excessive influence of the contract deadline. For instance, the two types of trees (CART and Chaid) that we have built predict that 0.90 of customers whose deadline is in April is at risk. If we consider that the variable target was built gathering data of February, the customers whose term is in April and have to regularly unsubscribe within the 60 days allowed, must become EXIT in February. Therefore, despite their good predictive capability, these models are useless for marketing actions, as a very simple model based on customer's deadlines will perform as well. The use of new methods is therefore necessary to obtain a predictive tool which is able to consider the fact that churn can be observed in different time periods, that is, ordered in calendar time.

The previous consideration gives the first reason why we decided to look for novel and different methodologies to predict churn.

We now turn our attention towards the application of the presented methodologies for modelling survival risk and to estimate the Lifetime value. In our case study the risk concerns the value that derives from the loss of a customer. The objective is to determine which combination of covariates affect the risk function, studying specifically the characteristics and the relation with the probability of survival for each customer. Our analysis is performed on the whole data sample.

3.5.1 The available data

The data set contains for each row a customer (ID) and each column represents a covariate (contract duration, sex, age,...). In particular, the data available for our analysis contains information that can affect the distribution of the event time, such as demographic variables, variables about the contract, the payment, the contacts and geomarketing. The only binary response variable, used as a dependent variable to build predictive models, includes two different types of customers: those who during the survey are active and those, instead, who regularly cancelled their subscription (EXIT status).

We remark that, due to the different time nature of the withdrawal, SUSPENSION status customers, who have not paid the subscription, although not cancelled, cannot be simply included in a classical analysis but, rather, require a specific treatment,

as in the survival analysis context. We remark that the target variable has been observed 3 months after the extraction of the data set used for the model implementation phase, in order to verify correctly the effectiveness and predictive power of the models themselves. We have available 606 variables and a sample of 3500 observations (customers), extracted from the company database.

Concerning explanatory variables, the available variables employed were taken from different databases used inside the company, which contained, respectively: socio-demographic information about the customers; information about their contractual situation and about its changes in time; information about contacting the customers (through the call centre, promotion campaigns, etc) and, finally, geo-marketing information (divided into census, municipalities and larger geographical sections information).

The variables regarding customers contain demographic information (age, gender, marital status, location, number of children, job and degree) and other information about customer descriptive characteristics: hobbies, pc possession at home, address changes.

The variables regarding the contract contain information about its chronology (signing date and starting date, time left before expiration date), its value (fees and options) at the beginning and at the end of the survey period, about equipments needed to use services (if they are rented, leased or purchased by the customer) and binary variables which indicate if the customer has already had an active, cancelled or suspended contract. There are also information about invoicing (invoice amount compared to different period of time - 2, 4, 8, 12 months).

The variables regarding payment conditions include information about the type of payment of the monthly subscription (postal bulletin, account charge, credit card), as well as other info about the changes of the type of payment. The data set used for the analysis also includes variables which give info about the type of the services bought, about the purchased options, and about specific ad-hoc purchases, such as number and total amount of specific purchases during the last month and the last 2 months.

The variables regarding contacts with the customer contain information about any type of contact between the customer and the company (mostly through calls to the call centre). They include many types of calling categories (and relatives sub-

categories). They also include information about the number of questions made by every customer and temporal information, such as the number of calls made during the last month, the last two months and so on.
Finally, geomarketing variables are present at large, and a great amount of work has involved their pre-processing and definition.
Regardless of their provenience, all variables have gone through a pre-processing feature selection step aimed at reducing their very large number (equal to 606). Such step has been performed using a combination of wrapping and filter techniques, going from dimensionality reduction to association measure ranking.

3.5.2 Classical Survival analysis

In order to build a survival analysis model, we have constructed two variables: one variable of status (that distinguishes between active and non active customers) and one of duration (indicator of customer seniority) . The first step in the analysis of survival data consists in a plot of the survival function and of the hazard.
We now consider the application of the Kaplan Meier estimator to our data. Figure 3.1 shows the estimated survival function. From Figure 3.1 note that the survival

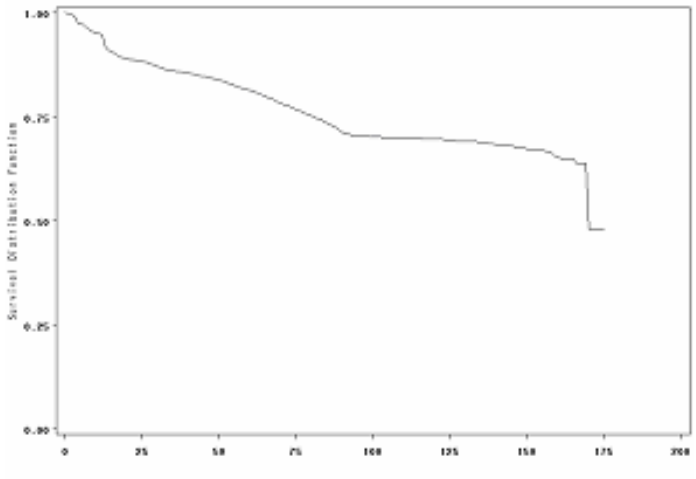

Figure 3.1: Descriptive Survival function

function has varying slopes, corresponding to different periods. When the curve decreases rapidly we have time periods with high churn rates; when the curve de-

creases softly we have periods of 'loyalty'. We remark that the final jump is due to a distortion caused by a few data, in the tail of the lifecycle distribution.
In Figure 3.2 we show the hazard function, that shows how the instantaneous risk rate varies in time. From Figure 3.2 we note two peaks, corresponding to months

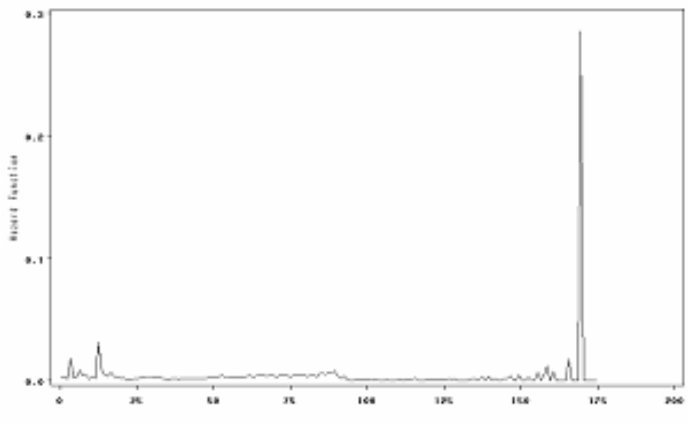

Figure 3.2: Hazard function

4 and 12, the most risky ones. Note that the risk rate is otherwise kept almost constant along the lifecycle. Of course there is a peak in the end, corresponding to what observed in Figure 3.1.
A very useful information, in business terms, is the calculation of the life expectancy of the customers. This can be obtained as a sum over all observed event times:

$$\sum_{j=1}^{T} \hat{S}(t_j) \times (t_j - t_{j-1}), \tag{3.49}$$

where $\hat{S}(t_j)$ is the estimate of the survival function at the j-th event time, obtained using the Kaplan Meier method, and t is a duration indicator. We remark that life expectancy tends to be underestimated if most observed event types are censored (i.e., no more observable).
We now move to the building of a full predictive model. We have chosen to implement first the classical Cox's model. The number of variables available is 606. The result, following a stepwise model selection procedure, is a set of about twenty explanatory variables. Such variables can be grouped in three main categories, according to the sign of their association with the churn rate, represented by the hazard ratio:

- variables that show a positive association (e.g. wealth of the geographic regions, the quality of the call center service, the sales channel)
- variables that show a negative association (e.g. number of technical problems, cost of service bought and payment method)
- variables that have no association (e.g. equipment rental cost, age of customer, number of family components).

To better interpret the previous associations we have considered the values of the hazard ratio under different covariate values. For example, for the variable indicating number of technical problems we have compared the hazard function for those that have called at least once with those that have not made such calls. As the resulting ratio turns out to be equal to 0.849, the risk of becoming churner is lower for "callers" than for "non callers".
A very important remark is that Cox model generates survival functions that are adjusted for covariate values. More precisely, the survival function is computed according to the following:

$$S(t, X) = S_0(t)exp(\sum_{i=1}^{p} \beta_i X_i), \tag{3.50}$$

Figure 3.3 shows a comparison between the survival curve obtained without covariates and the same curve adjusted for the presence of covariates. Figure 3.3 shows

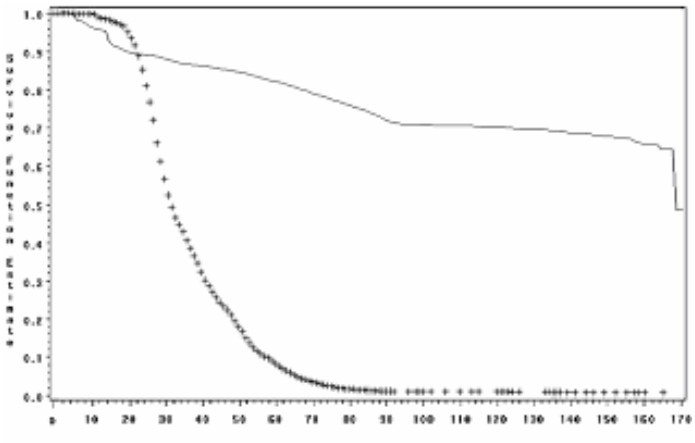

Figure 3.3: Comparison between survival functions

that covariates affect considerably survival times: up to two years of lifetime, the Cox survival curve (described by the symbols '+') is greater with respect to the baseline (described by the continuous curve). After such period the survival probability declines abruptly and turns out to be much lower for the remaining lifespan. We remark that, once a Cox model has been fitted, it is advisable to produce diagnostic statistics, based on the analysis of residuals, to verify if the hypotheses underlying the model are correct. In our case they were found to be correct, so we could proceed with predictive modelling. A further advantage of the survival analysis approach lies in its immediate translation in terms of lifetime value analysis, as we shall see in the next subsection.

3.5.3 Estimation of customer lifetime value

We now employ the results from survival analysis modelling to create models that allow to estimate the lifetime value of each customer, or, in a perhaps more useful aggregated analysis, for each segment of customers. In other words, survival analysis is useful to quantify, in precise monetary terms, how much is gained or how much is lost by moving through different strata corresponding to different survival curves. For instance ,how much is gained/lost if 0.08 of the clients, say, switch from buying service A to buying service B. Or, similarly, the relative gains when a certain percentage of clients change method of payment (e.g. moving between banking account, credit card and postal order).

In order to quantify gains and losses, a simple measure is to calculate the area between the two corresponding survival curves, as shown in Figure 3.4 below. Suppose the two survival curves correspond to two different services bought, say black and grey, corresponding to the colours of the two curves. In order to determine exactly the area in Figure 3.4 we need to specify a temporal period ahead, e.g. 13. In Figure 3.4, the difference between survival probabilities after 13 months of life of the customers (e.g. 13 months since the first contact), is equal to 0.078. This value should be multiplied by the difference in business margin between the two methods of payment, as given, for example, by the difference in costs. Such costs can be described by a gain table as in Table 3.1. From Table 3.1, a value of A is the relative gain if the client switches from PO to CC and, similarly, B and C corresponds to relative gains switching from PO to RID, and CC to RID where PO = postal order,

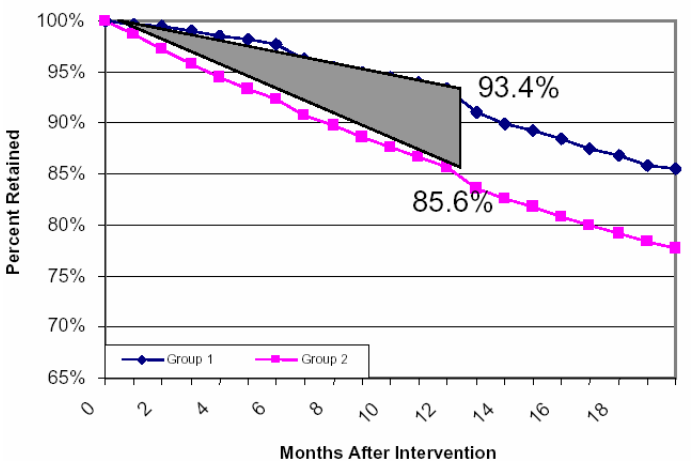

Figure 3.4: Evaluation of gain/losses by comparing survival curves

	PO	CC	RID
PO		A	B
CC			C
RID			

Table 3.1: Relative gains between different methods of payment

CC= payment through credit card, RID= payment through banking account.
In terms of Figure 3.4, if we assume that we start with an acquired client base of 1000 customers in both categories (product black buyers and product grey buyers), the results say that, after 13 months we will remain with 934 black and 856 grey. If the finance department tell us that product black is worth 10 euros and product grey 20 euros we have that, after 13 months, we lose 660 euros for black churners and 2880 for grey churners. In other words, the priority of the marketing department should be to build targeted campaigns for grey product clients. From a different perspective, if black and grey correspond to two different selling channels of the same product, or to two different geographical areas, it is clear that the black channel (or area) is much wiser in terms of customer retention. Often promotional campaigns are conducted looking only at increasing the customer base. Our results show that the number of captured clients should be traded with their survival or, better, lifetime value profile.

3.6 Application of Bayesian lifetime value models

3.6.1 Two step models lifetime value

We now proceed with Bayesian modelling, in a two step context. Firstly we apply the two step method explained in Section 3.1 to select variables. We have written a new R function to evaluate, for each covariate, a measure of importance. We report the results in Table 3.2 for the most important variables. It is possible to see that the most important variables to explain churn, are about information on disconnection, decoder rental , payment method, promotions, sale channel and contact with the call center.
After feature selection we used BUGS to implement a Bayesian counting process model. Given the model assumptions, this program performs the Gibbs sampler by simulating from the full conditional distribution. Table 3.3 shows the results. In particular for each covariate selected by our Bayesian feature selection approach we calculate, for each parameter, the mean , the standard deviation, the Markov Chain Monte Carlo error, the median and the Bayesian confidence interval. We have estimated our models with different MCMC chains. The most stable result is with 10000 iterations and 500 iteration as a burn-in. We have also used the idea

$Variable$	$p(y\|g)$	$p(g\|y)$
β info disconnection	0.2451	0.0472
β_2 decoder sold	0.2452	0.0472
β_3 decoder rental	0.2466	0.0475
β_4 payment credit card	0.2497	0.0481
β_5 promotion	0.2514	0.0484
β_6 channel of sell	0.2588	0.0491
β_7 ex decoder rental	0.2521	0.0488
β_8 special offers	0.2835	0.0546

Table 3.2: Two step model: the most important variables

$Variable$	$Mean$	Sd	$MCerror$	0.25	$Median$	0.975
β	0.7769	0.2123	0.00139	0.3547	0.7831	1.162
β_2	-1.632	2.223	0.08101	-5.938	-1.688	3.186
β_3	-1.731	0.6359	0.0308	-2.991	-1.718	-0.4818
β_4	-2.203	0.8412	0.04174	-3.715	-2.25	-0.3793
β_5	-1.368	0.6166	0.02468	-2.514	-1.396	-0.1127
β_6	-0.7287	1.626	0.09111	-3.206	-0.9579	3.382
β_7	-1.494	0.6678	0.02963	-2.845	-1.48	-0.215
β_8	0.67	2.141	0.1202	-3.957	0.6207	4.817

Table 3.3: Two step model: parameter estimation from the Bayesian Cox Model

of parallel multiple chains to check the convergence of the Gibbs sampler, following Gelman and Rubin (1992b). In particular, to generate the Gibbs posterior samples, we have used three parallel chains. Monitoring convergence of the chains, has been done via the Brooks and Gelman (1998) convergence-diagnostic-graph. Inspection of the Brooks and Gelman's diagnostic graphs for the most important covariates (Figure 3.5 and 3.6), show that BGR (Brooks and Gelman Ratio) converges to one, this show that convergence is archived for the two most important variables. We remark that this result is archived for all covariates in the model.

For each of the 3 chains BUGS depicts estimated parameters as a function of the iteration number. As is well known the BUGS software offers also a graph of the

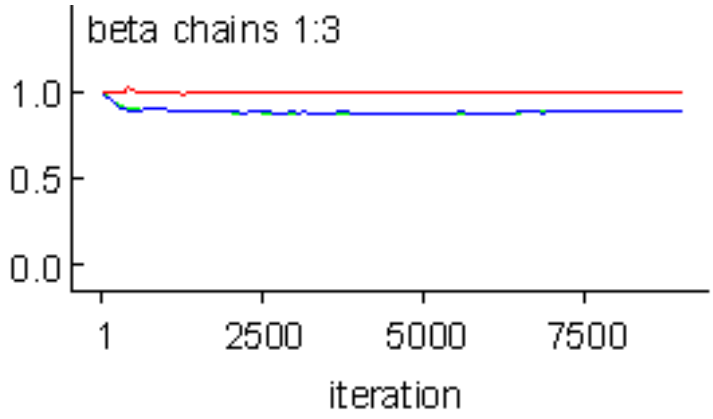

Figure 3.5: BGR for information on disconnection

autocorrelation function (ACF); the autocorrelation plot illustrates dependence between subsequent simulated observations. In our case, the ACF indicate fairly rapid mixing and thus good convergence of the parameter space with a reasonably small number of iterations. They are suppressed in this talk for lack of space. We also remark that for the model in Table 3.3, the estimated correlations between parameters that are quite low.

In order to compare classical and Bayesian Cox models we have developed model comparison. The results are shown in Table 3.4. In Table 3.4 the first column is the variable, the second and the third are the estimated mean and standard deviation in the Bayesian model and the last two columns are relative to the classical estimation for each variable, reporting the parameter MLE, the standard deviation and the

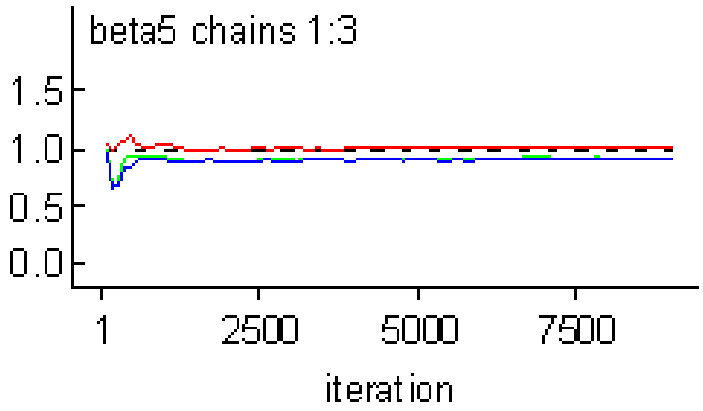

Figure 3.6: BGR for method of payment

$Variable$	$Mean$	Sd	$Estimate$	Sd	$p-value$
β	0.7769	0.2123	0.9396	0.2052	0.0001
β_2	-1.632	2.223	-1.4215	0.2647	0.0001
β_3	-1.731	0.6359	0.1164	0.1159	**0.3155**
β_4	-2.203	0.8412	0.2396	0.1356	**0.0772**
β_5	-1.368	0.6166	-0.8086	0.1510	0.0001
β_6	-0.7287	1.626	1.9636	0.1748	0.0001
β_7	-1.494	0.6678	-0.3392	0.1542	**0.0278**
β_8	0.67	2.141	1.0876	0.1206	0.0001

Table 3.4: Comparison of estimates from classical and Bayesian Cox model

p-value. As it is possible to see from the p-value, the variables β_3, β_4 and β_7 are equal to zero. The parameters β_3, β_4 and β_6 have different estimation in the two approaches. The BIC for the Bayesian Cox model is equal to 6583.411 and for the classical semi-parametric Cox model is equal to 6165.974. In particular, in Table 3.4, the variance of the estimates is quite high in the Bayesian model. To demonstrate the consistency of our proposed method for feature selection we next compare the previous results with the results from the one step approach.

3.6.2 One step models

We now study the application of Bayesian Model Averaging to our dataset. For computational reasons we have preselected 25 covariates which correspond to those that would be selected in a classical model selection step. We remark that six variables selected with the method described before are included.
First we use a full set of 25 covariates and after we compare the feature selection obtained from BMA with our proposed approach. We recall that there are 2^{25} possible models; we fit all the models and averaged over them to get parameter estimates and posterior probabilities of the parameters. Table 3.5 shows the top 3 models. Note that such models include 0.90 of the overall posterior probability, so that no much information is lost by reducing the model space. Table 3.5 shows the Bayesian model averaging computation for the covariates selected by our approach. From Table 3.6 we can see the posterior probabilities for each model and the number of selected variables for each model. From Table 3.7 we can see the posterior probabilities for each model and the number of selected variables for each model after the feature selection process. It can be seen that the estimation of the parameters is quite stable, as it does not vary too much with the variable selection procedure. To improve the previous results in the following parts of the talk we focus on more evoluted and more realistic Cox models, on which we shall apply our proposal BMA procedure. The results from the application of such models will be presented in the next chapter.

$Variable$	p	EV	$Model_1$	$Model_2$	$Model_3$
info activation	100	1.0783	1.1152	1.0793	1.0826
info administrative	100	1.5323	1.5274	1.5343	1.5134
β **info disconnection**	100	0.8512	0.8640	0.8596	0.8703
technical problem	100	-0.5071	-0.5159	-0.5098	-0.5078
β_5 **promotion**	100	-0.8985	-0.8963	-0.8920	-0.8749
β_6 **channel of sell**	100	1.6203	1.6415	1.6243	1.6220
β_4 **payment with credit card**	100	-0.6285	-0.6356	-0.6223	-0.6435
geographical area	100	0.3976	0.3991	0.3899	0.4154
β_8 **special offers**	100	3.1730	3.1294	3.1790	3.1676
β_7 **ex decoder rental**	100	-2.1571	-2.7982	-1.6625	-2.8616
β_3 **decoder rental**	50.1	-0.6230	-1.3120	.	-1.3717
β_2 **decoder sold**	59.7	-0.3894	-0.9544	.	-1.0077
β_i **...**	...	...	...	.	...

Table 3.5: One step model: Bayesian Model averaging results

$Model$	$PosteriorProbability$	$nVar$
Model_1	0.299	18
Model_2	0.166	16
$Model_3$	0.150	17
$Model_4$	0.148	16
$Model_5$	0.139	15

Table 3.6: One step model: the best 5 models

$Model$	$PosteriorProbability$	$nVar$
$Model_1$	0.456	5
$Model_2$	0.395	6
$Model_3$	0.071	6
$Model_4$	0.054	7
$Model_5$	0.023	6

Table 3.7: One step model: the best 5 models after feature selection

Chapter 4

Stratified lifetime value models

The purpose of this chapter is to test the hypothesis whether identical regression models are appropriate for different groups, that is, whether the relationships between the independent variables and survival are identical in different groups. To perform a stratified analysis, one must first fit the respective regression model separately within each group. In this chapter we propose a set of Bayesian methods and models to improve the results in the previous chapter. We shall propose Stratified One step models and Bayesian penalized models.

4.1 Improving Cox Models

In this Section we present a number of issues that can make the Cox model, more general and realistic. Such issues will be the basis, together with penalized methods, for our results on Bayesian Stratified models that will be presented in Section 4.5. Cox model has become the most used procedure for modelling the relationship of covariates to a survival or other censored outcomes (Therneau et al., 2000). Its form is flexible enough to allow time-dependent covariates as well as frailty terms and stratification. It has some restrictions. One of the restrictions to using the Cox model with time fixed covariates is its proportional hazards (PH) assumption, that is, that the hazard ratio between two sets of covariates has to be constant over time (this is due to the common baseline hazard function cancelling out in the ratio of the two hazards).

Thus, for fixed-time covariates, the exponent of a coefficient describes the relative change in the baseline hazard due to that covariate. The baseline hazard is typically

considered a 'nuisance parameter' and estimation of β is done by maximizing a profile likelihood with $\lambda_0(t)$ being substituted for an expression involving β and x, as well as the times at which failures occurred (Klein and Moeschberger, 1997). This expression is called the profile maximum likelihood estimate of $\lambda_0(t)$. The likelihood with $\lambda_0(t)$ 'profiled out' is called the partial likelihood by Cox (1972). For fixed-time covariates and independent observations, the partial likelihood is shown in equation 2.41. The value of β that that maximizes equation 2.41 is called the maximum partial likelihood estimate (MPLE). Typically in real applications it is not so easy to proof that the assumption for the Cox Model holds.

Now we review tools available to assess whether hazards can be considered proportional (PH assumption) across all covariates.

For binary covariates, as in our case, a comparison of nonparametric survival curve estimates may be sufficient to decide on PH because if the hazards were proportional, the survival curves for the two conditions would separate exponentially, and the two curves would not cross each other. Non-PH would imply that the relative risk changes over time for subjects who churn versus subjects who do not churn during the temporal period of study. For continuous covariates it is not sufficient to rely only on stratified survival estimates to assess PH because the choice of stratification points is subjective.

In this case an alternative is via the use of time-varying coefficients. That is, one or more coefficients multiplying their respective covariates varies with time. If the coefficient multiplying a covariate is not constant over time, then the impact of that covariate on the hazard varies over time, leading to non-PH. If PH holds, a plot of the coefficient versus time will be a horizontal line. Therefore, we can perform formal tests for specific forms of departure from PH. To illustrate formal tests of time-varying coefficients, we first describe the Schoenfeld residual, using the notation of Therneau et al. (2000).

Let $t_1, ..., t_d$ be the d unique ordered event times, and let $X_i(s)$ be the $p \times 1$ covariate vector for the $i-th$ individual at time s. For time-fixed covariates, this is just X_i. Also, define the 'weighted mean' of the $X_i(s)$ over those still at risk at time s as:

$$\overline{x}(\hat{\beta}, s) = \frac{\sum Y_i(s) exp(X_i(s)\hat{\beta}) X_i(s)}{\sum Y_i(s) exp(X_i(s)\hat{\beta})}, \tag{4.1}$$

where $Y_i(s)$ is is the predictable variation process indicating whether observation i is at risk at time s, so that $Y_i(s) = 1$ if observation i is still at risk at time s and is zero otherwise. The estimate $\hat{\beta}$ comes from fitting a Cox PH model. Then, a Schoenfeld residual is a $p \times 1$ vector that is defined at the $k-th$ event time as:

$$s_k = \int_{t_{k-1}}^{t_k} \sum_i \left[X_i(s) - \overline{x}(\hat{\beta}, s) \right] dN_i(s), \tag{4.2}$$

where $N_i(s)$ is a counting process that counts the number of events for observation i at time s. Thus, s_k sums the quantities $X_i(t_k) - \overline{x}(\hat{\beta}, t_k)$ over observations that have experienced the event by time t_k. With no tied event times, the $k-th$ Schoenfeld residual is the sum of contributions to the derivative of the log partial likelihood by subjects who have experienced events by t_k (Hosmer and Lemeshow, 1999).

A *scaled* Schoenfeld residual is equation 4.2 divided by an estimate of its standard deviation. Grambsch et al. (2000) show that the standard deviation is the square root of the weighted variance of $X_i(s)$ at time s. The scaled Schoenfeld residuals are used in a test of proportional hazards. For the $j-th$ covariate, Grambsch and Therneau (2000) express a time-varying coefficient as:

$$\beta_j(t) = \beta_j + \gamma_j g_j(t), \tag{4.3}$$

where $g_j(t)$ is a specific function of time. They show that the scaled Schoenfeld residuals have, for the j-th covariate, a mean at time t of approximately $\gamma_j g_j(t)$. Thus, a plot of the scaled Schoenfeld residuals by the event times may assess whether the coefficient γ_j is zero or not, and what the function $g_j(t)$ might be. A linear regression line can also be fitted to the plot along with a test for zero slope. A nonzero slope is evidence against PH.

As an alternative method of plotting, we can add the estimate of the regression coefficient to the scaled Schoenfeld residual to get a plot of the regression coefficient by time. We have done this, for our media company data, in Figure 4.1. Figure 4.1 shows scatter plots of the scaled Schoenfeld residuals by time for each single covariate from a classical Cox model. The smoothed curve in the plot is a natural spline with four degrees of freedom. The curve gives an indication of the path of the regression coefficient for that covariate by time. Ninety-five percent confidence bands are also given by dotted lines, using the variance of the estimated spline curve (Grambsch et al., 2000). The horizontal line is the estimate of the coefficient from

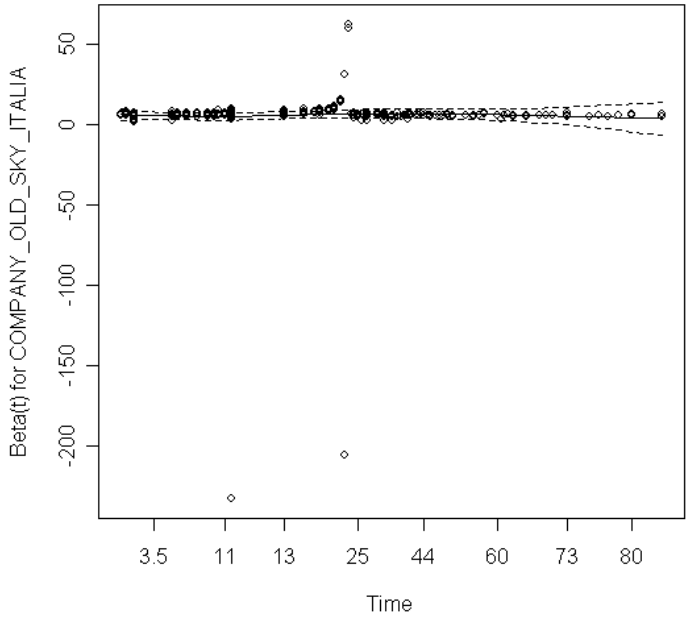

Figure 4.1: Shoenfeld residuals for company history

the Cox model. Figure 4.1 indicates a small changing effect on the hazard, and it is significantly linear at the 0.05 level. We can also use monotonic functions of time $g(t)$, such a $log(t)$, on the abscissa. Other specifications for g(t) lead to various tests for PH in the literature. See Grambsch et al. (2000) for details. We tried the transformation $g(t) = log(t)$.

Time transformations that are less influenced by outliers include rank time and the Kaplan-Meier (KM) transformation $g_j(t) = 1 - \hat{S}(t)$ where $\hat{S}(t)$ is the KM estimate (Grambsch et al., 2000). Alternatively, we might use a rank correlation test or just rely on the smoothed spline fits to the scatter plots as a way to visualize non-PH, especially if a nonlinear trend were suspected. There is also a limitation in the form of non-PH detected by scaled Schoenfeld residuals. Complicated forms of non-PH that involve interactions between covariates and time-dependent coefficients (e.g., a different coefficient function for each value or set of values of a covariate) cannot readily be detected unless we suspect them and construct a Schoenfeld plot for that subset of values. It is possible that evidence of time-varying coefficients appears because of other causes instead of non-PH.

Grambsch et al. (2000) list some of these reasons, including omitted covariates and incorrect functional forms for covariates. Omitted covariates are always a possibility and can cause non-PH (Grambsch et al., 2000). The addition of a frailty term may account for unmodeled covariates. We briefly discuss this issue later when we mention frailty models. For some other variables, as shown in Figure 4.2, there is much evidence on violation in PH. Other two tests that we can introduce are the

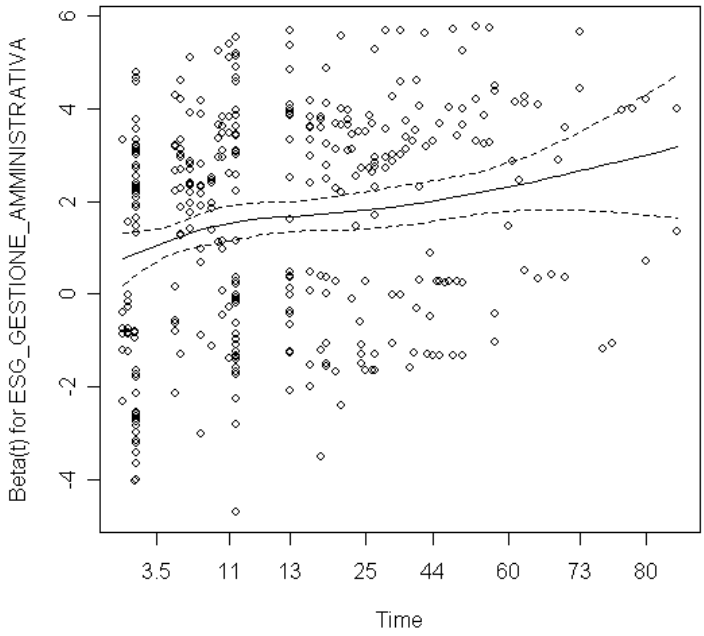

Figure 4.2: Shoenfeld residuals for administrative esigence

Andersen plot (Andersen et al. 1982, Klein and Moeschberger, 1997) and the Arjas plot. Based on the above assessments, we can conclude that, for our application the hazard rate may not be proportional over time across categories of some covariates. Specifically, the effects of covariates on the hazard rate may change over time.

There are several options for attempting to correct non-PH or to be used as alternatives to a PH model. An option is to use an accelerated failure time (AFT) model. Grambsch et al. (2000) show these models can be detected by the time-varying coefficient tests mentioned in this section.

AFT models are most appropriate in settings in which the time scale of the hazard function is either slower or faster (multiplicatively) than the time scale on which the measurements are made, as the covariates act by expanding or contracting time by

a factor $exp(X\beta)$.
Another alternative, is to stratify the model across levels of one or more covariates, leading to a stratified cox model, as will be shown in the next Section.

4.2 Stratified fixed effects Cox models

The Cox model can be extended to account for stratification. When a factor does not affect the hazard multiplicatively, stratification may be useful in model building. The strata divide the subjects into disjoint groups, each of which has a distinct (arbitrary) baseline hazard function but common values for the covariate dependent hazard (Grambsch et al., 2000). The hazard function for an individual i who belongs to stratum k is then

$$\lambda(t; x_i) = \lambda_k(t)exp(x_i'\beta), \tag{4.4}$$

Typically, strata are naturally defined within the context of the problem. For example, in medical research, multi-center clinical trials typically stratify on the clinic in which they are conducted (Grambsch et al., 2000).
The stratified Cox model allows a deviation from proportional hazards, and as such provides an alternative to the assumption of proportional hazards. The hazard functions for two different strata do not have to be proportional to one another. However, within a stratum, proportional hazards are assumed to hold. We take advantage of this use of stratification for our data.
The partial likelihood for stratified Cox models with K strata becomes a product of K terms, but where i ranges over only the subjects in stratum k, $k = 1, ..., K$.
Stratification entails fitting separate baseline hazard functions across strata. A baseline hazard function represents the hazard rate over time for an individual with all modelled covariates set to zero. With a stratified Cox model, a proportional hazards structure does not necessarily hold for the combined data, but is assumed to hold within each stratum. However, the coefficients on the included covariates are common across strata so that the relative effect of each predictor is the same across strata, unless there is a significant strata-by-covariate interaction, which means that the effect of the covariate differs within strata.
The estimated coefficients of a stratified Cox model are computed using the entire data set. One disadvantage of using a stratified model is that an effect of the strat-

ification covariate cannot be estimated in the model, at least in the usual sense of a coefficient estimate. This is a limitation if the stratification covariate is not merely a 'nuisance' variable that is recorded, but is of no substantive interest for the study (e.g., the clinic or hospital name at which recordings were made). However, if a model has been stratified on an important continuous variable that has been categorized, it is possible to also include the continuous variable in the model and thus estimate a relative effect for that covariate. The relative effect of the covariate is assumed to be the same within each stratum. In addition, the baseline hazard function within each stratum can also be estimated using; for example, Breslow's estimate.

For the stratified cox model it is possible to compute model diagnostics as for Proportional Hazard Cox Model, (see e.g.Grambsch et al., 2000). A formal test of overall goodness-of-fit for stratified Cox model was proposed by Parzen and Lipsitz (1999) and independently by May and Hosmer (1998). The test compares observed and (model-based) expected numbers of events within covariate risk groups and computes a chi-square test. The covariate regions are defined by predicted risk scores, $\hat{\psi}_i = exp(x_i'\hat{\beta})$, where $\hat{\beta}$ is the MPLE (Maximum Partial Likelihood Estimate) from the fit of a Cox model. The cut-points of, say, G regions are defined by percentiles of the $\hat{\psi}_i$ values, called percentiles of risk, such that each category ideally contains roughly the same number of observations. Each observation is classified into one of these G categories depending on its risk score, and (G-1) dummy variables are introduced into the Cox model. The score test of the resulting set of (G -1) coefficients constitutes a significance test for overall fit of the Cox model. Sample size guidelines given by Parzen and Lipsitz (1999) follow those for general chi-square tests: in order for the score test to reliably have an approximate chi-square distribution, at least 0.8 of the categories must have estimated expected count of at least five and all estimated expected counts should exceed one.

In our case, expected counts were estimated using the estimated asymptotic martingale residuals from the Cox model fit. The sum of the observed number of events minus the sum of the estimated martingale residuals within each category give the estimated expected count for that category (Parzen and Lipsitz, 1999) or (May and Hosmer, 1998). Parzen and Lipsitz (1999) mention in passing that their test can be used as a formal test for PH by using time intervals as well as risk groups to

divide the observations. However, in our experience this version of the test was very difficult to use correctly if we obeyed the sample size recommendations because it involved arbitrary decisions about the partitioning of the time-by-covariate space.

4.3 Bayesian stratified fixed effects Cox models

We now propose a Bayesian extension of the classical stratified Cox model as follow. We start with the Stratified Cox Model:

$$h_i(t) = h_{0i}(t)exp(\beta' x), \tag{4.5}$$

Stratum-specific baseline hazards, $h_{0i}(t)$ are assumed to be drawn from a Weibull family:

$$h_{0i}(t) = \rho_i t^{\rho_i - 1} exp(\rho_i \beta_{0i}),$$

We assume that the parameters of the model are: β_1,β_0, ρ_i where β_{0i} has a prior $N(\mu_0, \sigma_0^2)$, μ_0 is a flat prior and σ_0^2 an Inverse Gamma with specific value of parameter (e.g. 3,0.5).

We remark that ρ_i is a unit-specific shape parameter. In particular, if $\rho_i > 1$ there is an increasing hazard, $\rho_i \sim Ga(\alpha, \alpha^{-1})$, with $\alpha \sim Ga(c, d)$. In our case we have chosen c=3, d=10. The posterior distribution function of the coefficients can be derived via Gibbs Sampling.

4.4 Efficiency of Partial Likelihood methods

An important question that has not been discussed so far concerns the efficiency of the partial likelihood procedures. If for example, the data arise from a Weibull hazard function for T, given x, of the form $h(t|x) = \lambda\alpha t^{\alpha-1}e^{x\beta}$, then β can be estimated by standard parametric methods, as can the underlying hazard function $h_0(t) = \lambda\alpha t^{\alpha-1}$.

Therefore, if the partial likelihood is used to estimate β, some loss of information results from neglecting a portion of the available data. The crucial question is whether this loss of information is substantial. The issue of efficiency of the partial likelihood procedures has been examined by several authors: it appears that the efficiency is high when the underlying model truly obeys the proportional hazard

assumption and $|\beta|$ is not too large. This is an important result, because the partial likelihood methods are more robust than methods based on a specific parametric model.

Investigation of the efficiency of partial likelihood methods for β in anything but simple situation is difficult and must be done by simulation to a large extent.

To exemplify the problem, suppose that the data arise from a model with continuous hazard function:

$$h(t|x) = h_0(t;\theta)e^{x\beta}, \tag{4.6}$$

where $h_0(t;\theta)$ is some specified underlying hazard function involving a vector θ of unknown parameters. Given a censored sample involving observations on n individuals, assume that the likelihood function is of the usual form:

$$L_1(\beta,\theta) = \prod_{i=1}^{n}\left[h_0(t_i;\theta)e^{x_i\beta}exp\left(-\int_0^{t_i} h_0(u;\theta)e^{x_i\beta}du\right)\right]^{\delta_i}$$
$$\times\left[exp\left(-\int_0^{t_i} h_0(u;\theta)e^{x_i\beta}du\right)\right]^{1-\delta_i}, \tag{4.7}$$

where t_i is a lifetime or censoring time and $\delta_i = 1$ or 0 according to whether t_i is the former or the latter. Let us denote the expected information matrix arising from $L_1(\beta,\theta)$ as I:

$$\mathcal{I} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

where $I_{11} = E\left\{-d^2 log\frac{L_1}{d\beta d\beta}\right\}$, $I_{12} = E\left\{-d^2 log\frac{L_1}{d\beta d\theta}\right\}$, $I_{21} = E\left\{-d^2 log\frac{L_1}{d\theta d\beta}\right\}$ and $I_{22} = E\left\{-d^2 log\frac{L_1}{d\theta d\theta}\right\}$.

Let $\hat{\beta}$ be the m.l.e of β obtained by maximizing $L_1(\beta,\theta)$; the asymptotic covariance matrix for $\hat{\beta}$ is the upper $p \times p$ principal sub-matrix of I^{-1}, which is $V_1 = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$.

The partial likelihood in this situation is:

$$L_2(\beta) = \prod_{i=1}^{K}\left(\frac{e^{x_i\beta}}{\sum_{l\in R_i} e^{x_l\beta}}\right), \tag{4.8}$$

where $t_1 < ... < t_k$ are the distinct observed lifetimes in the sample and R_i is the risk set at t_i. The expected information matrix based $L_2(\beta)$ is the $p \times p$ matrix:

$$I_2 = \left[E\left(\frac{-d^2 logL_2}{d\beta_r d\beta_t}\right)\right], \tag{4.9}$$

Under suitable conditions the estimate $\tilde{\beta}$ of β obtained by maximizing $L_2(\beta)$ is asymptotically normal, with covariance matrix $V_2 = I_2^{-1}$. The asymptotic efficiency of $\tilde{\beta}$ relative to $\hat{\beta}$ can be obtained by comparing V_1 and V_2.

Computation of V_2 is however very difficult, and calculation of V_1 can also be difficult unless the censoring mechanism is a simple one. Thus even asymptotic results are hard to obtain and it is thus necessary to resort to simulation. Efron (1977) manages to obtain some general results by comparing V_1 and V_2 in a large class of models and reaches some fairly general conclusions. Along with results for some specific models mentioned below, that gives at least a broad impression of the efficiency properties of the partial likelihood methods.

The main general results (Efron, 1977) is that for certain types of models the partial likelihood m.l.e. $\tilde{\beta}$ is asymptotically fully efficient. This result is of limited direct use, since most realistic models are only approximated by a model of the type that Efron considers, though it does suggest that the efficiency of the partial likelihood procedures will be high in many instances. Other results indicate this is true in a few fairly simple models. For instance, Kalbfleisch(1978), Hensler et al.(1977), Efron(1977), Oakes(1977) all consider the case of an underlying exponential distribution, where $h_0(t)$ is constant, as in Section 1. In this case, when there is a single regressor variable and no censoring, the partial likelihood procedures are asymptotically fully efficient.

When there is censoring the efficiency falls off somewhat, but the loss does not appear to be great. In particular, papers have investigated the two-sample problem and find that the partial likelihood method is still asymptotically fully efficient at $\beta = 0$ if there is a common censoring time distribution in the two populations but there is a loss of efficiency when censoring time distribution are different.

Similar efficiencies appear to be obtained with more than one regressor variable and under other common proportional hazards models, though this has not been very thoroughly investigated. Peace and Flora (1978) report results of a small simulation study involving four regressor variables and underlying exponential, Weibull and Gompertz distributions, with some censoring. They found the efficiency of the partial likelihood methods to be fairly high in all cases. A few additional simulation results for the two sample problem (Gehan and Thomas, 1969) indicate that under a Weibull model the situation is similar.

The partial evidence accumulated so far thus suggest that the partial likelihood methods for β have good efficiency in a range of situations. Two points that should be remembered are that these results assume that a proportional hazards models is appropriate and that the efficiency can drop off rapidly if $|\beta|$ is large.

4.5 Stratified random effects Cox Models

Oakes (1992) reviews several models that have been proposed for multiple-event time data, including Andersen and Gill (1982). But none of these models allow for heterogeneity among individuals nor do these methodologies apply to data with ties. These disadvantages can be avoided by modelling the stratum effects (i.e. effects due to events occurring for the same subject) as independent random variables, themselves drawn from the same distribution, which may have to be estimated itself. This idea of frailty, which was introduced by Vaupel, Manton, and Stallard (1979), is particularly natural in the context of proportional intensity models. In some situations the extra random frailty component of the proportional intensity model is required only to get a correct inference on fixed effects of explanatory variables, whereas in other cases the distribution of the random subject effect could be one of the major interests. Now we would like to give attention to more sophisticated stratified models, which include random effects in a stratified Cox model. Again we shall consider frailty models from a Bayesian viewpoint. In real studies, there may be unmeasured factors. This suggests that the addition of subject-specific frailty terms into the partial likelihood may help the fit of a stratified Cox model.

A frailty is an unobserved continuous random variable that describes excess risk or 'frailty' for distinct groups, such as families or even single individuals, in addition to that described by measured covariates (Grambsch et al., 2000). Thus, frailties are like unobserved covariates (random effects). Individuals with greater frailty are expected to experience the event earlier than those with lower frailties. A Cox model with subject-specific frailties is a special case of a shared frailty model (Hougaard, 2000). The term shared comes from the use of such models to account for dependence among certain observations.

In a frailty model, the hazard conditional on the frailty is:

$$\lambda(t; X, \varpi) = \lambda_0(t) exp(X\beta + Z\omega), \tag{4.10}$$

where ϖ is a random variable, the frailty, Z is a vector of indicators that selects the appropriate ω term for each subject.

Note that the model that we have presented in Chapter 4 is a special case of the present one. For many frailty distributions, proportional hazards are only assumed conditionally. The marginal hazard formulation resulting from integrating the associated survival distribution with respect to the frailty distribution does not usually reflect proportional hazards (Hougaard, 2000).

Using frailties in Cox models is quite common. In fact, it is believed by some (Hougaard, 1995) that all models should contain frailties. When censoring is present, the bias is reduced.

To use frailties in our model, we must specify a distribution for the frailties; that is, the distribution from which the frailties are assumed to be a random sample. There are many conventional choices for the frailty distribution in the literature. Hougaard (1995) reviews many of these choices. We considered only two choices: a gamma distribution and a lognormal distribution. Both of these frailty distributions allow us to estimate parameters by maximizing a penalized partial log likelihood with penalty function equal to the log likelihood for a random sample of ω from the appropriate distribution (Grambsch et al., 2000). The parameters to be estimated are the coefficients in the ordinary Cox partial likelihood, plus any unknown parameters in the frailty distribution.

The log-likelihood given for the gamma frailty model is the log partial unpenalized likelihood integrated with respect to the frailty distribution. A likelihood ratio test (LRT) that the frailty variance exceeds zero is given by twice the difference between this integrated log likelihood and the log likelihood of a model without frailties. This statistic has an approximate chi-squared distribution with one degree of freedom (Grambsch et al., 2000). Estimation of the frailty variance θ is done in an outer loop of a Newton-Raphson algorithm for estimating β and ω. Assuming a fixed θ, $\hat{\beta}$ and $\hat{\omega}$ are found. Then, $\hat{\omega}$ is found by maximizing the profile likelihood with β and ω profiled out, and $\hat{\beta}$ and $\hat{\omega}$ are then re-estimated. Full details of the estimation procedures can be found in Grambsch et al. (2000). The standard errors for the coefficient estimates come from the inverse of the second derivative matrix of the penalized log likelihood. Because these estimates are computed assuming that the parameter θ is fixed, they are typically underestimated. The standard errors can

thus be corrected using the bootstrap procedure.

Although stratified Cox models are very popular for survival data, it is not the only flexible model available. The Cox model with time-fixed covariates assumes a multiplicative effect of covariates on the baseline hazard (except if covariates enter through stratification). Alternatively, the Aalen et al. (1980) additive hazard models consider the hazard as an additive combination of covariate terms, where the coefficients in the linear combination may be dependent on time, allowing the covariate effects to vary over time. Thus, covariates have an additive effect on the baseline hazard. This model measures additional excess risk due to the effects of a covariate in absolute terms rather then in relative terms (Klein and Moeschberger, 1997). In addition, frailty distributions can be nonparametric instead of having a form that is dependent on only a few parameters (Ibrahim et al., 2001). More generally the model assumption that allow classical estimation hold. In the next section we propose a new Bayesian approach to improve classical Cox results, based on stratified random effects models. To do this we need to introduce the issue of penalized likelihood for stratified models.

4.6 Penalized likelihood methods

In this section we introduce penalized likelihood methods that will be used in Section 4.5 in our proposal of a Bayesian Model Averaged random effect Cox model. The basic idea behind Penalized Maximum Likelihood Estimation is that there are two aims in estimation of a curve, g; one is to maximize fit to the data, as measured by the log likelihood $l(g)$, and the other is to avoid curves which exhibit too much roughness or rapid variation. Roughness of a curve g can be measured by a *roughness functional*, $J(g)$ in various ways. A typical choice of $J(g)$ is $\int g''(t)^2 dt$, which will have a high value if g exhibits a large amount of local curvature and zero if g is a straight line. The possibly conflicting aims of obtaining a high value of $l(g)$ while guarding against excessive values of $J(g)$ are reconciled by subtracting from the log likelihood a multiple of $R(g)$, called a *roughness penalty*, to obtain:

$$l_p(g) = l(g) - \lambda R(g). \tag{4.11}$$

A Penalized Maximum Likelihood Estimate is a curve g which maximizes the penalized log likelihood over the class of all curves satisfying sufficient regularity conditions

for $J(g)$ to be defined. The *smoothing parameter* λ controls the trade-off between high likelihood and smoothness and hence determines implicitly how much the data are smoothed to produce the estimate.

In particular, penalized likelihood is discussed in two rather different settings: function estimation and model choice. Penalized maximum likelihood estimation in nonparametric regression and density estimation were reviewed by Silverman (1985).In statistical inference about infinite-dimensional objects such as functions of one or more continuous variables, the principle of maximum likelihood is usually inadequate as a basis for estimation. The function in question may be, for example, a probability density or hazard function, a regression relationship, or the intensity function of a point process, so problems of this kind occur very widely indeed.

In parametric approaches to inference, the inherent impossibility of identifying an infinite-dimensional object from a finite amount of data is avoided, of course, by assuming that the function has a particular parametric form (a normal density, a quadratic regression curve, etc.), so that inference about only a finite number of parameters is involved. Very occasionally, there are other ways round the difficulties posed by dimension, as in the partial likelihood approach to Cox's proportional hazards model. Generally, however, it is necessary to confront the nonparametric nature of the problem explicitly, and penalized likelihood provides a natural and unifying approach for doing so.

The failure of maximum likelihood as a principle for estimation of functions really derives from a substantial inadequacy in the formulation of the inferential problem. No matter how successfully the likelihood function captures the relationship between the unknown function and the data, it cannot fully represent all that is known about the function, for it is impossible to imagine a scenario in which no other information is available, however informally, from the context.

Bayesian inference provides one approach to model explicitly such prior information; penalized likelihood provides an alternative, non-Bayesian, paradigm with a similar aim. There are formal connections between the two approaches.

Suppose we write the likelihood of the data Y given an unknown g as:

$$p(Y|g) \propto e^{L(g)}, \tag{4.12}$$

and place a prior distribution on g of the form,

$$p(g) \propto e^{-\frac{1}{2}\lambda J(g)}, \tag{4.13}$$

then by Bayes Theorem,

$$p(g|Y) \propto e^{L(g)-\frac{1}{2}\lambda J(g)}, \tag{4.14}$$

so that the maximum penalized likelihood estimator of g is identified as the mode of its posterior distribution, or *maximum a posteriori* estimator (see e.g. Green and Silverman, 1994). This connection should not be surprising: the underlying notion and justification of penalized likelihood is very close to that of specifying prior belief or knowledge about g, albeit in a non-probabilistic form. The extent to which the connection is useful is limited by two factors:

- whether is plausible in probabilistic terms as an expression of prior belief, for roughness functional $J(g)$ of interest, and
- whether the *maximum a posteriori* criterion is appealing in the context.

Regarding the first point, there are certainly difficulties. It is commonly the case the $J(g)$ is invariant to the addition of a constant, or other transformations of g, so that $p(g)$ is an improper distribution. More precisely, is it *partially improper*; for example, conditioning on the values of g at a few distinct points yields a proper distribution. But further, because of the infinite dimensions involved, there are apparent paradoxes. For example, in the case of cubic smoothing splines, it turns out that the prior and posterior distributions of g are entirely concentrated outside the space of smooth functions over which the optimum (the cubic smoothing spline) is sought.

Regarding the second point, there are general grounds for suspicion about using the mode as a summary of the distribution of a high-dimensional object. These issues are all explored further in Green and Silverman (1994).

4.6.1 Penalized likelihood for fixed effect Cox models

Suppose we have n observations (t_i, δ_i, X_i), where t_i is the possibly censored survival time, δ_i is the censoring indicator and X_i is a row vector of covariates for individual i. In the Cox proportional hazard model the parameters are estimated by maximizing

the partial log-likelihood that can be written as:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left(X_i\beta - ln \left(\sum_{h \in R_i} exp(X_h\beta) \right) \right), \tag{4.15}$$

where R_i is the set of all individuals at risk at time t_i. The maximum likelihood estimate of β is denoted by b.

We now define the penalized partial log-likelihood as:

$$l^{\lambda}(\beta) = l(\beta) - \frac{1}{2}\lambda p(\beta), \tag{4.16}$$

where $p(\beta)$ is a penalty function and λ is a non-negative weight parameter, which now is considered fixed. The factor $\frac{1}{2}$ is introduced for mathematical convenience. The value of β that maximizes the penalized partial likelihood depends on λ and is denoted by b^{λ}. With only one categorical covariate with c categories, we write $X_i = (X_{i1}, ..., X_{ic})$, where $X_{ij} = I[X_i = j]$ is a dummy variable for category j with corresponding regression coefficient β_j. If the first category is considered as a baseline, we have $\beta_1 = 0$. If the categories of X are not ordered and, if in addition, we may assume that the effects of the various categories are not too different from the mean effect, a suitable penalty function is:

$$p_0(\beta) = \sum_{j=1}^{c} (\beta_j - \bar{\beta})^2. \tag{4.17}$$

This function penalizes β_j's that are farthest from the mean. Covariates with ordered categories form a special class to which we will refer as ordinal covariates. If X is an ordinal covariate, the difference between the regression coefficient of two adjacent categories is supposed not to be large. This leads to a penalty functions that penalize first or second-order differences of consecutive β_j's:

$$p_1(\beta) = \sum_{j=1}^{c-1} (\beta_{j+1} - \beta_j)^2, \tag{4.18}$$

or

$$p_2(\beta) = \sum_{j=2}^{c-1} (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2. \tag{4.19}$$

It is informative to consider the null space of a penalty function, that is, the set of coefficients β for which $p(\beta) = 0$. For p_0 and p_1 the null space only contains $\beta_1 = ... = \beta_c = 0$, while for p_2, all linear sequences of β_j's that imply a linear covariate effect remain unpenalized as well. If there exists a prior belief about the

behaviour of the regression coefficients, this can be reflected by an appropriate choice of the penalty function, especially its null space.

The extension to more covariates is straightforward. As an example we consider two ordinal covariates X_1 and X_2 with c_1 and c_2 categories and regression coefficients β_j, $(j = 1, ..., c_1)$ and γ_k, $(k = 1, ..., c_2)$, respectively. In an additive model the regression coefficient δ_{jk} of cell (j, k) can be obtained as $\beta_j + \gamma_k$. If there is interaction between X_1 and X_2, all cells must be considered as separate categories. In both cases the γ_{jk}'s can be penalized by:

$$p_3(\delta) = \sum_{j=1}^{c_1} \sum_{k=1}^{c_2} (\delta_{jk} - \bar{\delta_{jk}})^2, \tag{4.20}$$

where $\bar{\delta_{jk}}$ is the mean of the δ_{jk}'s of the bordering categories. Observe that coefficients of boundary cells are included in this function. They can be left out by summing from 2 to $c_1 - 1$ and $c_2 - 1$, respectively. The null space of p_3 is again only $\delta_{jk} = 0$ for j and k.

All four penalty function $p_0, ..., p_3$ can be written in the form $p(\beta) = \beta' A \beta$, with A a symmetric non-negative definite matrix. With $p(\beta)$ in this form the penalized log-likelihood can be written as:

$$l^\lambda(\beta) = l(\beta) - \frac{1}{2}\lambda \beta' A \beta. \tag{4.21}$$

For a given λ, the first and the second derivatives of the penalized log-likelihood with respect to β are given by:

$$\frac{\partial l^\lambda}{\partial \beta}(\beta) = U^\lambda(\beta) = U(\beta) - \lambda A \beta = \frac{\partial l}{\partial \beta}(\beta) - \lambda A \beta \tag{4.22}$$

and

$$-\frac{\partial^2 l^\lambda}{\partial \beta^2}(\beta) = H^\lambda(\beta) = H(\beta) + \lambda A = -\frac{\partial^2 l}{\partial \beta^2}(\beta) + \lambda A, \tag{4.23}$$

respectively.

A Newton - Raphson procedure can now be used to estimate the penalized regression coefficients b^λ.

A first-order Taylor expansion of U^λ around the unrestricted estimate b leads to the following approximation of b^λ:

$$b^\lambda = [H^\lambda(b)]^{-1} H(b) b. \tag{4.24}$$

Note that in the Normal linear model, this is precisely the Ridge estimate of Hoerl and Kennard (1970), if A is chosen as the identity matrix. Since b^λ is an intrinsically biased estimator, standard errors of b^λ are not very informative. Instead, we will report the square root of the diagonal elements of $[H^\lambda(b^\lambda)]^{-1}$ which give an impression of the stability of the penalized estimates. These quantities may be called pseudo-standard errors.

4.6.2 Cross-validated partial likelihood

In this section we determine the weight parameter, λ by using the predictive power of the model as a criterion. Predictive power is conceptually different from explained variation; while the latter measures the fit to the data from which the model was derived, the predictive power is a measure for the fit to future data. As there are no future data, we mimic the prediction process by cross-validation; every observation is left out once and predicted by using all other observations. In a Normal linear model with observations Y_i, the predictive power can be measured by the predicted sum of squares (for short, PRESS), which is a function of the weight parameter λ:

$$PRESS(\lambda) = \sum_{i=1}^{n} (Y_i - X_i b^\lambda_{(-i)})^2, \tag{4.25}$$

and here $b^\lambda_{(-i)}$ denotes the 'leave-one-out' regression coefficient, that is estimated when a single i is left out of the observations.

In a likelihood based (Cox) model, a generalization of PRESS leads to the partial cross-validated log-likelihood (see e.g. Verweij and Van Houwelingen, 1993):

$$CVL(\lambda) = \sum_{i=1}^{n} l_i(b^\lambda_{(-i)}), \tag{4.26}$$

where $l_i(\beta)$ is the contribution of the individual i to the log-likelihood, defined as $l(\beta) - l_{(-i)}(\beta)$, with $l_{(-i)}(\beta)$ denoting the unrestricted log-likelihood without individual i. The value of λ that maximizes $CVL(\lambda)$ is the optimal weight parameter to be used in the penalized partial log-likelihood. For a Cox model the expression for $l_i(\beta)$ can be derived as follows:

$$l_i(\beta) = ln \prod_{t_h < t_i} \left[\left(1 - \frac{exp(X_i\beta)}{\sum_{k \in R_h} exp(X_k\beta)} \right)^{d_h} \right] p_{ii}^{d_i}. \tag{4.27}$$

In particular, $l_i(\beta)$ the conditional probability that individual i dies at time t_h, given the individual is alive just before t_h. Observe that for $l_i(\beta)$ the product only

includes the observed failure times before t_i. Hence, $l_i(\beta)$ can be interpreted as the log-probability that individual i survives at all occasions before t_i and dies at t_i.
In practice we will use the following approximation to the CVL:

$$CVL(\lambda) = \sum_{i=1}^{n} l_i(b^\lambda) - c(\lambda), \tag{4.28}$$

with

$$c(\lambda) = tr\left([H^\lambda(b^\lambda)]^{-1} \sum_{i=1}^{n} U_i(b^\lambda) U_i(b^\lambda)^T \right) \tag{4.29}$$

a term that corrects the sum of the individual contributions. Here, $U_i(\beta)$ denotes the vector of derivatives of $l_i(\beta)$ with respect to β.
If the components of the log-likelihood are independent, as in linear and logistic model, the first term in the CVL equals $l(b^\lambda)$, the log-likelihood of the model evaluated at b^λ, while $c(\lambda)$ in expectation equals the effective dimension:

$$e(\lambda) = tr[H^\lambda(b^\lambda)]^{-1} H(b^\lambda). \tag{4.30}$$

Hence, with independent components, $CVL(\lambda)$ is approximately equal to a penalized version of Akaike's Information Criterion:

$$AIC(\lambda) = l(b^\lambda) - e(\lambda). \tag{4.31}$$

We remark that $AIC(0)$ is the original definition of AIC, that is the log-likelihood minus the dimension of the model.
We observe, in particular, that in the Cox model, the components $l_i(\beta)$ are dependent, which implies that the sum of the $l_i(\beta)$'s is not equal to $l(\beta)$ and $AIC(\lambda)$ does not approximate $CVL(\lambda)$. However, the differences ΔAIC and ΔCVL from the null model, without covariates, are approximately equal. Hence, $AIC(\lambda)$ can be useful as an alternative criterion for the determination of the weight parameter in penalized partial likelihood.

4.6.3 Penalized likelihood for random effect Cox models

If we think at a real application, the idea is that individuals have different frailties, and that those who are most frail will die earlier than the others. Aalen (1989) provides theoretical and practical motivation for frailty models by discussing the impact of heterogeneity on analyses, and by illustrating how random effects can deal with it.

Frailty are usually viewed as unobserved covariates; this has led to the use of the EM algorithm as an estimation tool (see e.g. McGilchrist, 1983). However, the algorithm is slow, variance estimates require further computation, and no implementation has appeared in any of the more widely available packages.

Penalized models provide an alternative approach that, besides, can be conciliated within a Bayesian framework. The frailty terms are treated as additional regression coefficients which are constrained by a penalty function added to the log-likelihood. They are computationally similar to other shrinkage methods for penalized regression such as ridge regression, the lasso (see e.g. Hastie et al., 2001) and smoothing splines.

Algorithms for fitting Cox semi-parametric and parametric models can be extended to include penalty functions. These methods usually converge quickly and produce both point and variance estimates for model parameters.

We discuss below the link between penalized estimation and frailty models. In particular, we would like to show that if the frailty has a gamma distribution, then the shared frailty model can be written exactly as a penalized likelihood. We also show that Gaussian frailty models are closely linked to penalized models. We then turn to computational issues in implementing penalized techniques for fitting proportional hazard frailty models.

Assume that the data for subject i, who is a member of the $j-th$ of q families, follows a proportional hazards shared frailty model. The hazard can be written as:

$$\lambda_i(t) = \lambda_0(t)\varpi_{j(i)}e^{X_i\beta}, \tag{4.32}$$

where $j(i)$ denotes that individual i belongs to family j, $\varpi_{j(i)} = \varpi_j$ is the frailty for family j, X is the covariate matrix of dimension n by p, and β is a vector of regression coefficients. The ϖ_j's are independent and identically distributed from some positive scale family with density function $f(\varpi;\theta)$, having mean 1 and variance θ for identifiably.

If the ϖ's are known, the complete log-likelihood is:

$$\begin{aligned}&\textstyle\sum_{i=1}^{n}[\int_0^\infty Y_i(t)\left[log(\lambda_0(t)) + log(\varpi_{j(i)}) + X_i\beta\right]dN_i(t)\\ &\quad - \textstyle\int_0^\infty Y_i(t)\varpi_{j(i)}exp(X_i\beta)\lambda_0(t)dt + logf(\varpi_{j(i)};\theta)].\end{aligned} \tag{4.33}$$

If the ϖ is viewed as missing data, the problem can be approached using the EM algorithm. Let $\phi = \phi(s,\theta)$ be the Laplace transform of the distribution of ϖ,

and let $\phi^{(n)}(s)$ be its $n-th$ derivative with respect to s. Let $A_j = A_j(\beta, \lambda_0) = \sum \int_0^\infty Y_i(s) exp(X_i\beta) d\Lambda_0(s)$, where the sum is over the members of family j, and let d_j be the number of events in the $j-th$ family. The log-likelihood of the observed data:

$$L_m(\beta, \lambda_0, \theta) = \sum_{i=1}^{n} \delta_i log\left(\int_0^\infty Y_i(t) e^{X_i\beta} \lambda_0(t)\right) + \sum_{j=1}^{q} log\left[(-1)^{d_j} \phi^{(d_j)}(A_j)\right], \quad (4.34)$$

is found by integrating over the distribution of ϖ. For any fixed value of θ, the literature suggests maximizing the likelihood for β and λ_0 by an EM algorithm, which alternates between the following steps.

- M-step: treat the current estimate of ϖ as a fixed value or $offset$, and update β and update β and λ_0 as in usual Cox regression. Note that for given β and ϖ,
$$d\hat{\Lambda}_0(t, \beta, \varpi) = \frac{\sum dN_i(t)}{\sum Y_i(t) \varpi_{j(i)} exp(X_i\beta)}. \quad (4.35)$$
- S-step: Compute ϖ as the expected value given the current values β and λ_0 and the data.
$$\varpi_j = -\frac{\phi^{(d_j+1)}(\hat{A}_j)}{\phi^{(d_j)}(\hat{A}_j)}, \quad (4.36)$$
where $\hat{A}_j = A_j(\beta, \hat{\lambda_0}(\beta, \omega))$.

Previous equations requires the shared frailty model and unfortunately do not hold for more complex models. The literature (1997) suggests that estimation of θ be done by maximizing the profile log-likelihood:

$$L_m(\theta) = L_m(\hat{\beta}(\theta), \hat{\lambda_0}(\theta), \theta). \quad (4.37)$$

Although ϖ is not an explicit parameter of the observed log-likelihood, the EM algorithm provides an estimate of this vector.
The penalized regression formulation for the shared frailty model is most easily developed alternative version of the hazard:

$$\lambda_i(t) = \lambda_0(t) eX_i\beta + Z_i\omega, \quad (4.38)$$

which is equivalent to equation 4.33. In this case, $\varpi_j = exp(\omega_j)$, Z is matrix of q indicator variables such that $Z_{ij} = 1$ when subject i is a member of family j and

0 otherwise, and each individual belongs to only one family. Estimation under this model is done by maximizing a penalized partial log-likelihood

$$PPL = PL(\beta, \omega; data) - g(\omega; \theta), \tag{4.39}$$

over both β and ω. Here PL is the log of the usual Cox partial likelihood,

$$PL(\beta, \omega) = \sum_{i=1}^{n} \int_0^\infty \left[Y_i(t)(X_i\beta + Z_i\omega) - log \sum_k Y_k(t) exp(X_k\beta + Z_k\omega) \right] dN_i(t) \tag{4.40}$$

and g is a penalty function chosen by the investigator to restrict the value of ω. The parameter θ is a tuning constant which may be pre-specified or adapted to the data. Typically, one would choose the penalty function to 'shrink' ω toward zero and use θ to control the amount of shrinkage.

To estimate β and ω, one solves the score equations. Because the penalty function does not involve β, $\frac{\partial PPL}{\partial \beta} = \frac{\partial PL}{\partial \beta}$. Therefore, the score equations for β are identical to those for an ordinary Cox model treating $Z\omega$ as an offset term. If we define:

$$\bar{z}_j(t) = \bar{z}_j(\beta, \omega, t) = \frac{\sum Z_{ij} Y_i(s) exp[X_i\beta + Z_i\omega]}{\sum Y_i(s) exp[X_i\beta + Z_i\omega]}, \tag{4.41}$$

then,

$$\frac{\partial PPL}{\partial \omega_j} = \sum_{i=1}^{n} \int_0^\infty (Z_{ij} - \bar{z}_j(t)) dN_i(t) - \frac{\partial g(\omega; \theta)}{\partial \omega_j}. \tag{4.42}$$

We recall that for given β and ω, the Breslow estimator of the underlying hazard is:

$$d\hat{\Lambda}_0(t, \beta, \omega) = \frac{\sum dN_i(t)}{\sum Y_i(t) exp(X_i\beta + Z_i\omega)}, \tag{4.43}$$

which is just Equation 4.34 in different notation.

Let $\hat{\lambda}_i = \hat{\lambda}_i(\beta, \omega) = \int_0^\infty Y_i(s) d\hat{\Lambda}_0(t, \beta, \omega)$. Some algebra shows that the score equation for ω_j is:

$$\frac{\partial PPL}{\partial \omega_j} = \sum_{i=1}^{n} [Z_{ij}\delta_t - Z_{ij}\hat{\lambda}_i e^{X_i\beta + Z_i\omega}] - \frac{\partial g(\omega; \theta)}{\partial \omega_j} = 0. \tag{4.44}$$

Because of the structure of the matrix Z this equation simplifies to:

$$\frac{\partial PPL}{\partial \omega_j} = [d_j - \hat{A}_j e^{\omega_j}] - \frac{\partial g(\omega; \theta)}{\partial \omega_j} = 0, \tag{4.45}$$

where d_j and $\hat{A}_j$ are defined above.

The penalized likelihood can be fitted with the Newton-Raphson algorithm. In addition to the score vectors $\frac{\partial PPL}{\partial \beta}$ and $\frac{\partial PPL}{\partial \omega}$, this requires the Hessian of the penalized

partial log-likelihood:

$$H = H(\beta, \omega) = I+ = \begin{pmatrix} 0 & 0 \\ 0 & g'' \end{pmatrix},$$

where $I = I(\beta, \omega)$ is the usual Cox model information matrix, or the second derivative matrix of PL with respect to β and ω.

4.6.4 Penalized likelihood for gamma frailty models

Details of the EM approach for the shared gamma frailty model can be found in Nielsen et al. 1992. Here we demonstrate that for any fixed θ, the penalized log-likelihood with appropriate choice of penalty function and the observed-data log-likelihood have the same solution. Let the frailty have a gamma distribution with mean 1 and variance $\theta = 1/\nu$. The density of ϖ can be written as:

$$log\left[f(\varpi;\nu)\right] = (\nu-1)log(\varpi) - \nu\varpi + \nu log(\nu) - log\Gamma(\nu). \tag{4.46}$$

This has a Laplace transform of $\phi(s) = (1+s/\nu)^{-\nu}$. The derivatives of $\phi(s)$ are:

$$\phi(s)^{(d)} = \left(-\frac{1}{\nu}\right)^d \left(1+\frac{s}{\nu}\right)^{-(\nu+d)} \prod_{i=0}^{d-1}(\nu+i), \tag{4.47}$$

and reduces to:

$$e^{\omega_i} = \frac{d_j+\nu}{\hat{A}_j+\nu}. \tag{4.48}$$

There exists in the literature the following Lemma. The solution to the penalized partial likelihood model, with penalty function :

$$g(\omega,\theta) = -1/\theta \sum_{j=1}^{q} \left[\omega_j - exp(\omega_j)\right], \tag{4.49}$$

coincides with the EM solution for any fixed value of θ.

To have an idea of the proof we consider for β the EM and penalized methods have the same score equation, which includes $Z\omega$ as a fixed offset. Thus if the solutions for ω be the same, those for β will be also. Let $(\hat{\beta}, \hat{\omega})$ be a solution to the the EM process. Then $\hat{\omega}$ must satisfy the previous equation exactly, not just as an update step. Rearranging terms, we see that:

$$\hat{A}_j = exp(-\hat{\omega}_j)(d_j+\nu) - \nu. \tag{4.50}$$

Substituting this into the penalized score equation and simplifying with $\nu = \frac{1}{\theta}$ a fixed quantity, we see that:

$$\frac{\partial PPL(\hat{\beta}, \hat{\omega})}{\partial \hat{\omega}_j} = [d_j - \hat{A}_j e^{\hat{\omega}_j}] - \frac{\partial g(\hat{\omega}; \theta)}{\partial \hat{\omega}_j} \tag{4.51}$$

and with some algebrical replacement the previous equation is equal at:

$$[d_j - e^{\hat{\omega}_j}(d_j + \frac{1}{\theta} - \frac{1}{\theta}e^{\hat{\omega}_j})e^{\hat{\omega}_j}] + \frac{1}{\theta}(1 - e^{\hat{\omega}_j}) = 0. \tag{4.52}$$

This shows that the solution to the EM algorithm is also a solution to the penalized score equations.

Therefore, for any fixed θ the penalized log-likelihood and the observed-data log-likelihood have the same solution, although these two equations are not equal to one another. Furthermore, if we let $PPL(\theta) = PPL(\hat{\beta}(\theta), \hat{\omega}(\theta), \theta)$, then we can write the profile log-likelihood for θ, as $PPL(\theta)$ plus a correction that only involves θ and d_j. Using the fact that each row of Z has exactly one 1 and $q-1$, we see that the Cox PL for $(\hat{\beta}, \hat{\omega})$ must be the same as that for $(\hat{\beta}, \hat{\omega} + c)$ for any constant c. Some algebra shows that the value of c which minimizes the penalty portion of the PPL is such that:

$$\sum_{i=1}^{q} e^{\hat{\omega}_j} = q. \tag{4.53}$$

Using the identities, recalling that they hold only at the solution point, we can show the following:

$$L_m(\theta) = PPL(\theta) + \sum_{j=1}^{q} \nu - (\nu + d_j)log(\nu + d_j) + \nu log\nu + log\left(\frac{\Gamma(\nu + d_j)}{\Gamma(\nu)}\right). \tag{4.54}$$

Shortly we give an idea about the proof to justify the previous equation. In particular we obtain the realized value of the marginal log-likelihood at the solution point in terms of the penalized likelihood for the shared frailty model. Expanding gives:

$$\begin{aligned} L_m(\beta, \lambda_0; \theta) &= \sum_{i=1}^{n} \delta_i \log\left(\int Y_i(t) e^{X_i\beta} d\Lambda_0(t)\right) \\ &+ \sum_{j=1}^{q} \left[-d_j \log\nu - (\nu + d_j)\log(1 + \frac{A_j}{\nu}) + \log\frac{\Gamma(\nu + d_j)}{\Gamma(\nu)}\right]. \end{aligned} \tag{4.55}$$

The log-profile likelihood for θ is just this function restricted to the one-dimensional curve defined by the maximizing values of $\hat{\beta}(\theta)$, $\hat{\omega}(\theta)$, $\hat{\lambda}_0(\theta)$. On that curve $\hat{A}_j = (d_j + \nu - \nu e^{\hat{\omega}_j})/e^{\hat{\omega}_j})$. With this substitution, after some rearrangement we get:

$$\begin{aligned} L_m(\theta) &= \sum_{i=1}^{n} \delta_i \log(\hat{\lambda}_i e^{X_i\hat{\beta} + Z_i\hat{\omega}}] + \sum_{j=1}^{q}[-(\nu + d_j)\log(\nu + d_j) \\ &+ \nu\log(\nu e^{\hat{\omega}_j}) + \log\Gamma(\nu + d_j) - \log\Gamma(\nu)], \end{aligned} \tag{4.56}$$

where δ_i is a $= 0/1$ indicator for an event for individual i. Subtracting and adding the penalty function $g(\omega;\theta) = -\frac{1}{\theta}\sum_{j=1}^{q}\omega_j - exp(\omega_j)$ evaluated at $\hat{\omega}$ results in:

$$\begin{aligned} L_m(\theta) &= \sum_{i=1}^{n}\delta_i \log(\hat{\lambda}_i e^{X_i\hat{\beta}}) - g(\hat{\omega};\theta) \\ &+ \sum_{j=1}^{q}[-\nu\hat{\omega}_j + \nu e^{\hat{\omega}_j} - (\nu+d_j)\log(\nu+d_j) \\ &+ \nu\log(\nu e^{\hat{\omega}_j}) + \log\Gamma(\nu+d_j) - \log\Gamma(\nu)]. \end{aligned} \tag{4.57}$$

The previous Equation can be express in the following form:

$$\Rightarrow PPL(\theta) + \sum_{j=1}^{q}\left[\nu - (\nu+d_j)\log(\nu+d_j) + \nu\log\nu + \log\left(\frac{\Gamma(\nu+d_j)}{\Gamma(\nu)}\right)\right]. \tag{4.58}$$

Note that, because considerable loss of accuracy can occur if one subtracts values of the log-gamma function, it is computationally advantageous to exploit the property of the Gamma function:

$$\log\left(\frac{\Gamma(\nu+d_j)}{\Gamma(\nu)}\right) = \sum_{i=0}^{d_j-1}\log\left(\frac{\nu+i}{\nu+d_j}\right), \tag{4.59}$$

rather then

$$\log\left(\frac{\Gamma(\nu+d_j)}{\Gamma(\nu)}\right) = \log(\Gamma(\nu+d_j)) - \log(\Gamma(\nu)). \tag{4.60}$$

It is useful to consider $L_m(\theta) + \sum_{j=1}^{q} d_j$, rather then $L_m(\theta)$, because the profile log-likelihood converges to $PL(\hat{\beta}) - \sum_j d_j$ as the variance of the random effect goes to zero. Adding $\sum_j d_j$, to $L_m(\theta)$ makes the maximized marginal likelihood from a frailty model with small θ comparable to the maximized likelihood from a non-frailty model.

The computational algorithm for a shared gamma frailty that we have developed consists of an inner and outer loop. For any fixed θ Newton-Raphson iteration is used to solve the penalized model in a few (usually 3-5) steps, and return the corresponding value of the PPL. The other loop chooses θ to maximize the profile likelihood, which is easily done as it is a unimodal function of one parameter.

All of the results presented in this section were dependent on the correct choice of a penalty function. For gamma frailties, we have shown that, the penalty function that links the penalized and EM results is directly related to the density of the random effect; the log of the density for ω, where $exp(\omega)$ has a gamma distribution, is equal to $\omega - exp(\omega)]/\theta$ plus additional terms not involving ω.

Similarly, it can be shown that the penalty needed for a Gaussian frailty is related to a log-density (see e.g. Thearneau et al. 2000).

4.7 Penalized Likelihood: Computational Issues

Thus far, we have discussed the relationship between frailty models and penalized likelihood estimation. In this section, we describe several issues important for our computational implementation of penalized likelihood methods for Cox models with random effects.

Consider a Cox model with both constrained and unconstrained effects. The model is fit by maximizing the penalized partial log-likelihood (PPL). We assume that θ is fixed. Consider the set of hypotheses $z = C(\beta', \omega')' = 0$, where $(\beta', \omega')'$ is the combined vector of $p + q$ parameters, and C is a $kxp + q$ matrix of full row rank k, $k \leq p + q$. Gray(1992) suggests that

$$V = H^{-1} I H^{-1}, \tag{4.61}$$

to be used as the covariance estimate of the parameter estimates. He recommends a Wald type test statistic, $z'(CH^{-1}C')^{-1}z$, with generalized degrees of freedom

$$df = trace[(CH^{-1}C')^{-1}(CVC')]. \tag{4.62}$$

The total degrees of freedom for the model ($C = I$) simplifies to

$$df = trace[HV] = trace[H(H^{-1}(H - G)H^{-1})] = (p + q) - trace[GH^{-1}]. \tag{4.63}$$

Under H_0, the distribution of the test statistic is asymptotically the same as $\sum_i e_i X_i^2$, where e_i are the k eigenvalues of the matrix $(CH^{-1}C')^{-1}(CVC')$ and X_i are i.i.d. standard Gaussian random variables. In non-penalized models, the test statistic has mean $\sum_i e_i$ and variance $2\sum_i e_i^2 < 2\sum_i e_i$, because $0 \leq e_i \leq 1$. Using a reference chi-square distribution with $df = \sum_i e_i$ the test will tend to be conservative.

Verweij and Van Houwelingen (1993) discuss penalized Cox models in the context of restricting parameter estimates. They use H^{-1} as a "pseudo standard error", and an "effective degrees of freedom". With this variance matrix, the test statistic $\acute{z}(CH^{-1}C')^{-1}z$ is an usual Wald test. To choose an optimal model they recommend either the Akaike Information Criterion (AIC) which uses the degrees of freedom described above.

In this paper we like to follow a new approach based on the cross-validated (partial) log-likelihood CVL, which uses a degrees of freedom estimate based on a robust variance estimator. In our case simulation experiments for the related problem of

penalized smoothing splines in Cox regression, based on nonparametric statistical approaches, suggest that this is the more reliable choice for tests. In our implementation, the computation of the degrees of freedom and variance matrices are specialized to avoid any intermediate steps that would give a q by q result, where q is the number of constrained coefficients.

When performing estimation with frailty models, memory and time considerations can become an issue. For instance, if there are 300 subjects, each with a frailty term, and 4 other variables, then the full information matrix has $304^2 = 92416$ elements. The Cholesky decomposition must be applied to this matrix within each Newton-Raphson iteration. In our R implementation, we have applied a technique that can provide significant savings in space and time. If we partition the information matrix of a Cox shared frailty model according to the rows of X and Z, and arrange the matrix as

$$I = \begin{pmatrix} I_{ZZ} & I_{ZX} \\ I_{XZ} & I_{XX} \end{pmatrix},$$

then the upper left corner will be a diagonally dominant matrix, having almost the form of the variance matrix for a multinomial distribution. Adding the penalty further increases the dominance of the diagonal. Therefore, using a *sparse* computation option, where only the diagonal of I_{ZZ} is retained, should not have a large impact on the estimation procedure.

Ignoring a piece of the full information matrix has a number of implications. First, the speed of the Cholesky factorization is increased dramatically. Second, the savings in space can be considerable. If we use the sparse option with the example above a savings of over 0.95 in memory space. Third, because the score vector and likelihood are not changed, the solution point is identical to the one obtained in the non-sparse case, discounting trivial differences due to distinct iteration paths. Fourth, the Newton-Raphson iteration may undergo a slight loss of efficiency so that l-2 more iterations are required. However, because each N-R iteration requires the Cholesky decomposition of the information matrix, the sparse problem is much faster per-iteration than the full matrix version. Finally, the full information matrix is a part of the formulas for the post-fit estimates of degrees of freedom and standard error.

In a small number of simple examples, the effect of the sparse approximation on

these estimates has been surprisingly small.

We have found two cases where our sparse method does not perform acceptably. The first is if the variance of the random effect is quite large and in this case, each N-R iteration may require a large number of iterations. The second is if one group contains a majority of the observations. The off diagonal terms are too important to ignore in this case, and the approximate K-R iteration does not converge.

4.8 Bayesian penalized likelihood

In this section we propose a new method to estimate the smoothing parameter λ that controls the trade-off between high likelihood and smoothness and hence determines implicitly how much the data are smoothed to produce the estimate. In order to obtain a non-parametric smooth hazard of survival function, it is quite easy to fit a frailty model and Cox proportional hazards model using a Penalized Likelihood on the hazard function. In this framework, left truncated and censored data are allowed. Clustered and recurrent survival times can be studied (the Andersen-Gill (1982) approach has been implemented for recurrent events) and an automatic choice of the smoothing parameter is possible using an approximated cross-validation procedure. Typically, when we build stratified models, the cross validation method is not implemented for two strata to choose the best smoothing parameter λ . There is only a possible solution in a very simple case, for a Cox proportional hazard model with no covariates. In order to solve our problem that considers many covariates and two strata variable, we propose a new methodology to choose the best smoothing parameter with a Bayesian kernel density estimation. The kernel method extends to the estimation of a density function in more than one dimension. As a descriptive exercise, a two-dimensional density estimate can be constructed by applying the first equation with a two-dimensional kernel function in the form:

$$\hat{f}(y_1, y_2) = \frac{1}{n} \sum_{i=1}^{n} w(y_1 - y_{1i}; \lambda_1) w(y_1 - y_{2i}; \lambda_2), \tag{4.64}$$

where $\{y_{1i} y_{2i}; i = 1, ..., n\}$ denote the data and (λ_1, λ_2) denote the joint smoothing parameters. We assume, in our case, that for each strata variable there is a positive smoothing parameter. It is possible also to study multivariate version of kernel function: several papers describes a variety of more sophisticated techniques for constructing and displaying density estimation that can be carried out in three, four

and more dimensions. In order to go beyond the exploratory and graphical stage it is necessary first to understand more about the behaviour of these estimators and to derive some basic properties. Although many theoretical results exist, simple expressions for means and variances of the estimators are enough to allow ideas of interval estimation and hypothesis testing to be discussed, and to motivate techniques for choosing an appropriate bandwidth to employ with a particular dataset.

The mean of a density estimator can be written as:

$$E\left\{\hat{f}(y)\right\} = \int w(y-z;\lambda)f(z)dz, \tag{4.65}$$

This is a convolution of the true density function f with the kernel function w. Smoothing has therefore produced a biased estimator, whose mean is a smoothed version of the true density. A Taylor series expansion then produces the approximation:

$$E\left\{\hat{f}(y)\right\} \approx f(y) + \frac{\lambda^2}{2}\sigma_w^2 f"(y), \tag{4.66}$$

where σ_w^2 denotes the variance of the kernel function, namely $\int z^2 w(z)dz$.

Since $f"(y)$ measures the rate of curvature of the density function, this expresses the fact that $\hat{f}$ underestimates f at peaks in the true density and overestimates at troughs. The size of the bias is affected by the smoothing parameter λ. The component σ_w^2 will reduce to 1 if the kernel function w is chosen to have unit variance.

Through another Taylor series argument, the variance of the density estimate can be approximated by:

$$var\left\{\hat{f}(y)\right\} \approx \frac{1}{n\lambda} f(y)\alpha(w), \tag{4.67}$$

where $\alpha(w) = \int w^2(z)dz$. As ever, the variance is inversely proportional to sample size. In fact the term $n\lambda$ can be viewed as governing the local sample size, since λ controls the number of observations whose kernel weight contributes to the estimate at y. It is also useful to note that the variance is approximately proportional to the height of the true density function.

The combined effect of these properties is that, in order to produce an estimator which converges to the true density function f, it is necessary that both λ and $1/n\lambda$ decrease as the sample size increases. A suitable version of the central limit theorem can also be used to show that the distribution of the estimator is asymptotically normal.

A similar analysis enables approximate expressions to be derived for the mean and

variance of a density estimate in the multivariate case. In p dimension, with a kernel function defined as the product of univariate components w, and with smoothing parameters $(\lambda_1, ..., \lambda_p)$.There are results for more general kernel functions.
It is helpful to define an overall measure of how effective $\hat{f}$ is in estimating f. A simple choice for this is the *mean integrated squared error* (MISE) which, in the one-dimensional case, is:

$$MISE(\hat{f}) = E\left\{\int \left[\hat{f}(y) - f(y)\right]^2 dy\right\} \tag{4.68}$$

$$= \int \left[E\left\{\hat{f}(y)\right\} - f(y)\right]^2 dy + \int var\left\{\hat{f}(y)\right\} dy, \tag{4.69}$$

This combination of bias and variance, integrated over the sample space, has been the convenient focus of most of theoretical work carried out on these estimates. In particular, the Taylor series approximations described allow the mean integrated squared error to be approximated as:

$$MISE(\hat{f}) \approx \frac{1}{4}\lambda^4\sigma_w^4 \int f"(y)^2 dy + \frac{1}{n\lambda}\alpha(w), \tag{4.70}$$

establishing the properties of the estimators which employ variable bandwidths, is more complex. In order to construct a density estimate from the observed data it is necessary to choose a value for the smoothing parameter λ. An overall measure of the effectiveness of $\hat{f}$ in estimating f is provided by the mean integrated squared error. From the approximate expression given there it is straightforward to show that the value of λ which minimizes MISE in an asymptotic sense is:

$$h_{opt} = \left\{\frac{\gamma(w)}{\beta(f)n}\right\}^{1/5}, \tag{4.71}$$

where $\gamma(w) = \alpha(w)/\sigma_w^4$ and $\beta(f) = \int f"(y)^2 dy$. This optimal value for h cannot immediately be used in practice since it involves the unknown density function f. However, in our case, it is very informative in showing how smoothing parameters should decrease with sample size, namely proportionately to $n^{1/5}$, and quantifying the effect of the curvature of f through the factor $\beta(f)$.
These ideas involved in cross-validation. In the context of density estimation, Rudemo (1982) applied these ideas to the problem of bandwidth choice, through estimation of the integrated squared error (ISE).

$$\int \left\{\hat{f}(y) - f(y)\right\}^2 dy = \int \hat{f}(y)^2 dy - 2\int f(y)\hat{f}(y) dy + \int f(y)^2 dy, \tag{4.72}$$

The last term on the right hand side does not involve λ. The other terms can be estimated by:

$$\frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i}^{2}(y)dy - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(y_i), \tag{4.73}$$

where $\hat{f}_{-i}(y)$ denotes the estimator constructed from the data without the observation y_i. It is straightforward to show that the expectation of this expression is the MISE of $\hat{f}$ based on $(n-1)$ observations, omitting the $\int f^2$ term. The value of h that minimises this expression therefore provides an estimate of the optimal smoothing parameter.

Techniques known as *biased cross validation* and *smoothed cross validation* also aim to minimise ISE but use different estimates of this quantity. These approaches are also strongly related to the "plug in" approach. In our application we employed different methods with multivariate density estimates: Optimal smoothing, normal optimal smoothing, cross validation smoothing parameter. Jones et al. (1996) give a helpful and balanced discussion of methods of choosing the smoothing parameter in density estimation. We describe now our Bayesian proposal employed to choose λ based on stochastic simulation. We use methods of generation of quantities with discrete or continuous probability distribution. We know that a generic λ is a positive number defined in $[0, \infty)$. In this framework we use some basic results about generation of continuous quantities in the probability integral transform stating that if x is continuous with distribution function F, then $u = F(x) \sim U[0, 1]$. This results is easily shown as:

$$F_u(y) = P(u \leq y) = P(F(x) \leq y) = P(x \leq F^{-1}(y)) = F(F^{-1}(y)) = y, \tag{4.74}$$

if $0 < y < 1$. Exploring the fact that F has an inverse, one finds that $F^{-1}(u) \sim F$. Our first aim is to generate a sequence a numbers deterministically, which embody the features of randomness as far as possible. In general, we use a recurrence relation of the form:

$$X_n = f_n(X_{n-1}, ..., X_0), n = 1, 2, 3, ... \tag{4.75}$$

or more usually, $X_n = f_n(X_{n-1}, ..., X_{n-r})$ for some r. This is less computationally expensive than taking $r = n$. We call X_0 the seed of the sequence. The seed is very important, since it is this that determines the entire sequence. We start with a very simple example.

Suppose that we require a set of random integer observations, generated from a Uniform distribution in the range $[0, M-1]$, i.e., $P(X = i) = 1/M \forall i, i = 0, ..., M-1$. If we let $M = 100$, then we might define $f(x)$ to be the nearest integer to $(100 \times fractionalpartoflog_e)$. Then, if $X_0 = 82$, $log_e = 4.406$, so that $X_1 = 41$. Similarly, $X_2 = 71$ and $X_3 = 26$, but then we find that $X_i = 26$, $\forall i \geq 3$. Similarly, if we set $X_0 = 39$, then $X_24 = X_48 = ... = 39$, and we only have 24 different values. There are various problem with ad-hoc functions:

- They always lead to periodic sequences. Once a number is repeated, a cycle starts. Note that the period is always less then M.
- They rarely cover the underlying parameter space well. For example, only four of the cycle of 24 above are greater than 75, and none lie between 77 and 94.
- The choice of starting value may be crucial.

Therefore, the function f must be chosen very carefully, so as to ensure that these undesirable properties are minimized.

4.8.1 Empirical choice of the smoothing parameter

Most data analyses take the form of obtaining observations $x = (x_1, ..., x_n)$ from some unknown probability distribution F and constructing an estimate $\hat{\lambda}(x)$ for some parameter of interest λ, given this observations. Having obtained $\hat{\lambda}(x)$, we then wish to assess the accuracy as an estimate of the true value, λ. If $\hat{\lambda}$ is unbiased, then the most common measure of this accuracy is the standard error given by:

$$SE(\hat{\lambda}) = \sqrt{(Var_F(\hat{\lambda}))}. \tag{4.76}$$

For example if $\lambda = E_F(X)$, then

$$\hat{\lambda}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{4.77}$$

is an unbiased estimator for λ. Further, if

$$\sigma^2(F) = Var_F(X), \tag{4.78}$$

then

$$SE(\hat{\lambda}) = \left(\frac{\sigma^2(F)}{n} \right)^{\frac{1}{2}}. \tag{4.79}$$

In this case, $\sigma^2(F)$ is a function of the unknown distribution function F, but this can also be estimated from the data by:

$$\sigma^2(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{4.80}$$

So that an estimate of the $SE(\hat{\lambda})$ is given by:

$$\hat{SE}(\hat{\lambda}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}. \tag{4.81}$$

Thus, if $\theta = E_F(X)$, the formula for the standard error is quite simple and easily estimable. However, for the other estimators such as the sample correlation coefficient, such formulae do not exist. Traditionally the statistical literature has concentrated upon obtaining approximations to $SE(\hat{\lambda}$ for particular estimators, avoiding estimators which do not have such formulae. However, over the past significant advances have been made in this area, firstly by the introduction of the Jacknife estimate error, and then by Bootstrap.

4.8.2 Jackknife and Bootstrap methods to choose the smoothing parameter

The Jacknife was introduced as a method for reducing the bias of a serial coefficient estimator, based upon splitting the sample into two half sample. The original method was subsequently generalized by splitting the sample into g groups of size h, with $g = nh$. The method of the Jacknife was then based upon this generalized method, method with $g = n$ and $h = 1$, and avoided the, often analytically complex, equation to compute $SE(\hat{\lambda}$ altogether, going directly to a generalization of the $\hat{SE}(\hat{\lambda})$ as follows.

Let $x_{(i)}$ be the data set with the i^{th} datum removed so that:

$$x_{(i)} = \{x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n\}, \tag{4.82}$$

and let $\hat{\lambda}_{(i)}$ be the statistic $\hat{\lambda}$ evaluated for the data set $x_{(i)}$. Then the jacknife estimate of the standard error of $\hat{\lambda}$ is given by:

$$\hat{SE}_{jack}(\hat{\lambda}) = \left(\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\lambda}_{(i)} - \bar{\lambda_{(.)}})^2\right)^{\frac{1}{2}}, \tag{4.83}$$

where $\bar{\lambda_{(.)}} = \frac{\sum_{i=1}^{n}\hat{\lambda}_{(i)}}{n}$. As an alternative to the standard error, we might also be interested in the bias of our estimator, defined as $bias = E(\hat{\lambda}) - \lambda$. We can define the

Jackknife estimate of bias to be $bias_{jack} = (n-1)(E_{\hat{F}}(\hat{\lambda}) - \hat{\lambda})$. That is the difference between the expected value of the empirical distribution and the estimated value, so this is simply $(n-1)(\bar{\lambda_{(.)}} - \hat{\lambda})$. Once again, if $\hat{\lambda} = \bar{x}$, then the estimate of the bias is zero i.e., we have an unbiased estimator. The Jackknife method (for standard errors) can be thought of algorithmically as follows with the R software.

- Step 1: Set i=1
- Step 2: Remove the $i-TH$ observation and calculate $\hat{\lambda}_{(i)}$ from the remaining observations.
- Step 3: Increase i and if $i \leq n$, return to STEP 2, else continue to STEP 4.
- Step 4: Calculate $\hat{SE}_{jack}$.

We consider here a set of data and we are interested in the smoothing parameter. As we have described before, we can estimate the smoothing parameter for example, with cross-validation technique, but how accurate is this estimate? We can assess this using the Jackknife algorithm. We have from our algorithm an object which list the $\hat{\lambda}_{(i)}$ values. By subtracting $\hat{\theta}$, the mean, from each of this values we can see which output have the largest influence upon the smoothing parameter between the samples. In order to better understand the Jackknife results, we propose the Bootstrap. Efron defined an empirical estimate, $\hat{F}$, given by $\hat{F}(x) = \frac{i}{n}, x_{(i)} \leq x \leq x_{(i+1)}$. Then, $\sigma^2(\hat{F}) = \frac{i}{n}\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2$, where the x_i^* are independent, identically distributed random variables drawn from $\hat{F}$, i.e.: the data set X^* consists of n observations randomly sampled with replacement from $x_1, ..., x_n$. Then,

$$\hat{\sigma}_{boot} = \sqrt{Var_{\hat{F}(x)}\hat{\lambda}(X^*)}. \tag{4.84}$$

Therefore, if $\lambda = E_F(X)$, then:

$$\hat{SE}_{boot}(\hat{\lambda}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2} = \sqrt{\frac{n-1}{n}}\hat{SE}(\hat{\lambda}). \tag{4.85}$$

It can be shown that asymptotically, $\hat{SE}_{boot}$ agrees with $\hat{SE}_{jack}$ and further, that the jackknife is in fact a linear approximation to the boostrap.
Obviously we can gain an empirical estimate of $\hat{\sigma}_{boot}$ by randomly selecting a large number of bootstrap sample, $x_1^*, ..., x_B^*$, and calculating $\hat{\lambda}(x_i^*)$ for each. Thus we

obtain:

$$\hat{SE}_{boot}(\hat{\lambda}) = \sqrt{\frac{\sum_{b=1}^{B}\left(\hat{\lambda}(x_b^*) - \bar{\lambda}(.)\right)^2}{B-1}}. \tag{4.86}$$

This approximation to the true standard error can be made arbitrarily accurate by taking suitably large B. This is essential difference between the jackknife and the bootstrap, since we can take as many bootstrap samples as we like, but we are restricted to only n jackknife samples. There are two aspects to the bootstrap procedure: replacement of F by $\hat{F}$ to form estimates of functionals of interest. We have implement with R software the bootstrap in order to obtain a set of bootstrap replicates $\hat{\lambda}_i^*$, $i = 1, ..., B$ and the bootstrap estimate of the standard error. We compare the Jackknife and the Bootstrap estimates with the corresponding standard error. The standard error for the bootstrap estimate is given as 0.129 as opposed to the 0.143, given by the jackknife method earlier. In order to assess the estimation for the smoothing parameter we prefer the boostrap method. In this context suppose that we are in a one-sample situation, where the data are obtained by random sampling from an unknown distribution, F. Let $\hat{\lambda} = \hat{\lambda}(F)$ be the plug-in estimate of a parameter of interest, $\lambda = \lambda(F)$, and let $\hat{SE}$ denote some reasonable estimate of the standard error for $\hat{\lambda}$, based perhaps upon bootstrap or jackknife solutions. Under most circumstances, it turns out that for large sample sizes, the distribution of $\hat{\lambda}$ becomes approximately normal, with mean λ and variance approximated by $\hat{SE}$. Therefore, if we base our estimate upon a sample $x_1, ..., x_n$, then for large n,

$$Z = \frac{\hat{\lambda} - \lambda}{\hat{SE}} \sim N(0, 1). \tag{4.87}$$

This result suggests that for large n, a $100(1-\alpha)\%$ confidence interval for λ can be constructed as:

$$\left[\hat{\lambda} - Z_{(1-\frac{\alpha}{2})}\hat{SE}, \hat{\lambda} - Z_{\frac{\alpha}{2}}\hat{SE}\right]. \tag{4.88}$$

We call this the standard confidence interval for λ. However, this interval relies upon the assumption of normality. If this assumption cannot be justified, then we can use the bootstrap to construct non-parametric confidence intervals in two way. First we describe how to obtain bootstrap intervals, without resorting to assumptions of normality as discussed above. In essence, we estimate the distribution of the $Z-statistic$ directly from the data, and use this distributional estimate to construct our confidence intervals. We begin by generating B bootstrap samples, and then

computing the bootstrap version of Z for each. We then construct percentiles for the Z statistic and use these to construct our intervals.

More formally, we generate B bootstrap samples $x_1^*, ..., x_B^*$, and for each we compute:

$$Z^*(b) = \frac{\hat{\lambda}^*(b) - \lambda^*}{\hat{SE}}(b), \tag{4.89}$$

where $\hat{\lambda}^*(b)$ is the value of $\hat{\lambda}$ for the bootstrap sample x_b^* and $\hat{SE}^*(b)$ is the estimated standard error of $\hat{\lambda}^*$ for the bootstrap sample x_b^*. The $\alpha - th$ percentile of $Z^*(b)$ is then estimated by $\hat{Z}_\alpha$, such that:

$$\frac{1}{B}\sum_{b=1}^{B} I_{(Z^*(b) \leq \hat{Z}_\alpha)}(b) = \alpha. \tag{4.90}$$

For example, if $B = 1000$, the estimate of the 5% point is the $50 - th$ largest value of the $Z^*(b)$. Finally, the bootstrap $t - confidence$ interval is given by:

$$\left[\hat{\lambda} - \hat{Z}_{(1-\frac{\alpha}{2})}\hat{SE}, \hat{\lambda} - \hat{Z}_{\alpha/2}\hat{SE}\right]. \tag{4.91}$$

The second solution is based on the percentile interval that works as follows. If, as before, we generate B bootstrap samples, $x_1^*, ..., x_B^*$, and for each we compute $\hat{\lambda}^*(b)$, the estimate of λ based upon the bootstrap sample x_b^*, then the percentile interval for λ is defined as:

$$\left[\hat{\lambda}^*_{\alpha/2}, \hat{\lambda}_{(1-\alpha/2)}\right], \tag{4.92}$$

where $\hat{\lambda}^*_\alpha$ is the $100\alpha th$ percentile of the bootstrap distribution i.e., if $B = 1000$, then $\hat{\lambda}^*_{0.05}$ is simply the $B\alpha th = 50th$ largest of the $\hat{\lambda}^*$ values.

Before to start with our Bayesian penalized likelihood approach, we show here a short remark on a possible feature research on Bayesian bootstrap to analyze the quality for the estimation of smoothing parameter, λ. The Bayesian version simulates the posterior distribution of the parameter λ, rather than the estimated sampling distribution of $\hat{\lambda}$ (an estimate od λ).Each Bayesian bootstrap replication generates a posterior probability for each x_i, where the values of X that are not observed have zero posterior in the same way that they have zero probability under the sample cumulative distribution function. The posterior probability is centred at $1/n$, but has some degree of variability. We introduce this variability by generating $n - 1$ random variables from $U(0, 1)$ distribution, $u_1, ..., u_{n-1}$, for each replication. Then we let:

$$g_i = u_{(i)} - u_{(i-1)}, i = 1, ..., n \tag{4.93}$$

where $u_{(i)}$ is the i^{th} smallest u_i, with $u_{(0)} = 0$ and $u_{(n)} = 1$. Then $g = (g_1, ..., g_n)$ is the vector of probabilities associated with the data values $(x_1, ..., x_n)$ for that replication. Repeated replications will then give an example of the Bayesian bootstrap distribution of X, and therefore of any parameter of this distribution.

For example, if $\lambda = E_F(X)$, then at each replication, we calculate the mean of X as if g_i were the probability that $X = x_i$, i.e.: we calculate $\sum_{i=1}^{n} g_i x_i$. The distribution of the values of $\sum_{i=1}^{n} g_i x_i$ over all Bayesian bootstrap replications gives the Bayesian boostrap distribution of the mean of X.

in practice, the bootstrap and the Bayesian bootstrap differ only in how the probabilities are attached to the x_i. However, the interpretations of the resulting distributions are different. The Bayesian bootstrap generates likelihood statements about parameters rather then frequency statements about statistics under assumed values for parameters, and thus has an inherent advantage over the bootstrap.

4.8.3 Bayesian choice of the smoothing parameter

Given a sample y, from a distribution with joint probability distribution $f(y|\lambda)$ and a prior for λ given by $p(\lambda)$, the Theorem of Bayes relates the posterior $\pi(\lambda|y)$ to the prior via the formula:

$$\pi(\lambda|y) \propto L(y|\lambda)p(\lambda), \tag{4.94}$$

where the constant of proportionality is given by:

$$\left(\int L(y|\lambda)p(\lambda)d\lambda \right)^{-1}. \tag{4.95}$$

Given the posterior, and in the case where $\lambda = (\phi, \psi)$ is multivariate, for example, we may be interested in the marginal posterior distribution, such as:

$$\pi(\phi|y) = \int \lambda = (\phi, \psi|y)d\psi. \tag{4.96}$$

Alternatively, we might be interested in summary inferences in the form of posterior expectations, for example,

$$E_\pi(g(\lambda)) = \int g(\lambda)\pi(\lambda|y)d\lambda. \tag{4.97}$$

Often, explicit evaluation of these integrals is not possible and, traditionally, we would be forced to use numerical integration or analytic approximation techniques. Markov Chain Monte Carlo provides an alternative whereby we sample from the

posterior directly, and obtain sample estimates of the quantities of interest.

Suppose that we have some distribution,$\pi(x)$, $x \in E \subseteq R^d$, which is known only up to some multiplicative constant. We commonly refer to this as the target distribution. If π is sufficiently complex so that we cannot sample from it directly, an indirect method for obtaining samples from π is to construct a Markov chain with state space E, and whose stationarity (or invariant) distribution is $\pi(x)$. Then, if we run the chain for long enough, simulated values from the chain can be treated as a sample from the target distribution and used as a basis for summarising important features of π.

Under certain regularity conditions, the Markov chain sample path mimics a random sample from π. Given realizations $X^t : t = 0, 1, 2, ...$ from such a chain, typical asymptotic results include the distributional convergence, when $t \to \infty$ of the realizations i.e., $X^t \xrightarrow{d} \pi(x)$ and the consistency of the ergodic average, for any scalar functional λ, i.e.,

$$\frac{1}{n}\sum_{t=1}^{n} \lambda(X^t) \overset{n\to\infty}{\to} E_\pi[\lambda(X)], a.s. \tag{4.98}$$

There are many important implementation issues associated with MCMC methods. These include amongst others the choice of sampler, the number of independent replications to be run, the choice of starting values, how long to run the samplers, and both estimation and efficiency problems.

In our case, we are interested in Markov chain defined on continuous state space, in general. A Markov chain is generated by sampling the new state of the chain, based only upon information regarding the current state i.e., at time t, we generate the new state of the chain from a density dependent only upon x^t:

$$x^{t+1} \sim K(x^t, x), (= K(x|x^t)). \tag{4.99}$$

We call K the transition kernel for the chain, and is uniquely describes the dynamics of the chain. Under certain conditions (the chain is both aperiodic and irreducible), the Markov chain will converge to its stationarity distribution i.e.,

$$P(X^t \in A) \to \pi(A), \forall A \in E. \tag{4.100}$$

Assuming that a stationary distribution exists, it is unique if the chain is irreducible. Irreducibility is a condition on the chain, which essentially means that any set of states can be reached from any other set of states, within a finite number of moves.

We shall come back to this concept later. Suppose that we have an irreducible Markov chain with stationarity distribution, $\pi(x)$. Then the Ergodic Theorem states that:

$$\bar{f}_n = \frac{1}{n}\sum_{t=1}^{n} f(x^t) \rightarrow E_\pi(f(x)), n \rightarrow \infty. \tag{4.101}$$

We call $\bar{f}_n$ the ergodic average. Thus, assuming that the chain has stationary distribution $\pi(x)$, and is irreducible, we can use the Markov chain sample path to obtain an estimate of the value of any functional of the variable X. This means that, since many integrals in which we are interested can be thought of as expectations, we can form Monte Carlo approximations to these integrals by sampling via a Markov chain. This technique is called Markov Chain Monte Carlo.

The usual approach to Markov chain theory is to start with some transition kernel, determine conditions under which there exists an invariant or stationary distribution, and then to identify the form of that limiting distribution. MCMC methods involve the solution of the inverse of this problem whereby the stationarity distribution is known and it is the transition kernel that needs to be identified.

There are a number of methods for developing Markov chains with a given stationary distribution.

A more general way to construct MCMC samplers is as a form of generalised rejection sampling, where values are drown from approximate distributions and "corrected" in order that, asymptotically, they behave as random observations from the target distribution. This is the motivation for methods such as the Metropolis Hastings algorithm which sequentially draws candidate observations from a distribution, conditional only upon the last observation, thus inducing a Markov chain with transition density $K(x, y)$ and exhibiting detailed balance for π i.e.,

$$\pi(x)K(x, y) = \pi(y)K(y, x), \tag{4.102}$$

the chain has stationary density, $\pi(.)$. The method which we used to choose the Bayesian smoothing parameter can be written as follows. We begin with a density for generating candidate observations. We allow this candidate generating density (or proposal density) to depend upon the current state of the chain, and we denote it by $q(x^t, y)$. In general, the introduced chain will not satisfy the reversibility condition, so we introduce an acceptance function $\alpha(x^t, y)$, and accept the candidate observation, so that $x^{t+1} = y$ with probability $\alpha(x^t, y)$. However, unlike rejection

sampling, if the candidate observation is rejected, the chain remains at x^t, so that $x^{t+1} = x^t$.

It can be shown that the optimal form for the acceptance function, in the sense that suitable candidates are rejected least often and computational efficiency is maximised, is given by:

$$\alpha(x^t, y) = min\left(1, \frac{\pi(y)q(y, x^t)}{\pi(x^t)q(x^t, y)}\right) \tag{4.103}$$

so that the transition kernel is given by:

$$P_H(x, A) = \int_A K_H(x, y)dy + r(x)I_A(x), \tag{4.104}$$

where

$$K_H(x, y) = q(x, y)\alpha(x, y), r(x) = 1 - \int_E q(x, y)\alpha(x, y)dy, \tag{4.105}$$

and K_H satisfies the reversibility condition implying that the Kernel, P_H also preserves detailed balance for π. We need only simulate from q, which we can choose arbitrarily. Moreover, and this can be of crucial importance, we only need to know π up to proportionality, since any constants of proportionality cancel in the numerator and denominator of the calculation of α. The price for simplicity is that if q is poorly chosen, then the number of rejections can be high, so that the efficiency of the procedure can be low. In our case we build two different algorithms that considers two different proposal distribution. We remark that from the bootstrap procedure the maximum value for λ was 20. So we know that the λ is between $[0, < \infty]$ and we propose firstly a Normal distribution as proposal and at the end also a continuous Uniform distribution $[0, 30]$. When we use as candidate a Normal distribution, we note that this distribution is symmetric, i.e., $q(x, y) = q(y, x)$, the acceptance function reduces to:

$$\alpha(x, y) = min\left(1, \frac{\pi(y)}{\pi(x)}\right) \tag{4.106}$$

We remark also that the acceptance function is independent of σ in the Gaussian case, but the value of σ has a significant impact upon the acceptance rate of the chain. We show the results of this implementation in the application. Here we advise the user to identify several possible tuning parameters, note their defaults and look at the sensitivity of the results to varying them. In penalized likelihood, after the choice for the smoothing parameter, the regression estimated parameter are obtained

using the robust Marquardt algorithm (Marquardt, 1963) which is a combination between a Newton-Raphson algorithm and a steepest descent algorithm. When frailty parameter is small, numerical problems may arise. To solve this problem, an alternative formula of the penalized log-likelihood is used (see Rondeau et al., 2003 for further details). Cubic M-splines of order 4 are used for the hazard function, and I-splines (integrated M-splines) are used for the cumulative hazard function. In the next section we report also a list of problems that we can find in the set of classical procedures.

4.9 Application of Bayesian stratified models

The previous methodology will now be applied to a real case to estimate customer survival. First we show the results for Classical Stratified Cox Model. As we can see in Table 4.1, the first column is the variable selected by the classical stratified procedure, the second is the relative estimate for each variable, then the hazard ratio and finally the sign of the association between each variable and the target variable (a mixture of relationship duration and customer status). As we can observe, for some variables there are some computational problems for the estimation. In particular, for 'smart card offers' and 'old customer', the estimate and the hazard ratio are undefined. This because the algorithm does not always converge. In order to evaluate the importance of each variable, we have compute the relative AIC. The results are in Table 4.2. Table 4.2 defines an ordering for variable selection. The most important variables, selected by our proposal on Bayesian feature selection, are technical problems, smart card offers, geographical area, and some other. We now implement a new procedure based on an extension of Bayesian Model Averaging (Madigan et. al 1999). In particular, we propose a natural extension of the classical stratified Cox Model in a Bayesian paradigm.

The results are in Table 4.3 where p is the probability of inclusion for each variable across the models, EV is the expected value for each variable, that derive from Bayesian Model averaging, and finally for each model the parameter estimate. At the last rows of the table, we have estimates for each stratus variables. In particular the variance of estimates across models is lower than in the previous one step Bayesian Model Averaging. If we look at the results, the estimates across the models in Table

Variable	*Estimate*	*Hazardratio*	*Association*
info activation	1.018	2.77	+
β info disconnection	1.1577	3.18	+
technical problem	-0.4131	0.662	-
contractual variation	-0.1605	0.852	NA
Moovie package	0.1889	1.210	NA
β_8 special offers	3.2696	26.3	+
special discount offers	3.4625	31.9	+
info administrative	1.6123	5.01	+
payment methods	-0.1493	0.861	NA
β_5 promotion	-0.9596	0.383	-
Sport package	-0.086	0.917	NA
payment with bancomat	0.8063	2.24	+
geographical area	0.4246	1.53	+
smart card offers	19.184	∞	∞
old customer	6.1058	∞	∞
β_i	...	...	...

Table 4.1: Classical Stratified Cox Model: results

Variable	*AICcontribution*
technical problem	3274.4
smart card offers	3275.1
geographical area	3276.5
β_5 promotion	3277.6
info activation	3284.0
payment with bancomat	3316.9
β_8 special offers	3389.3
special discount offers	3389.8
old customer	3414.6
info administrative	3458
...	...

Table 4.2: Classical Stratified Cox model: the best models in terms of relative AIC contributions

4.3, are very similar. In our research this suggest some new theoretical field of research in order to improve this approach. Table 4.4 shows the best 5 models found by our Stratified fixed effects Cox Bayesian Model averaging. Table 4.4 presents for each model its posterior probability and its dimension based on the number of covariates. In Table 4.5, we present the results based on Stratified random effects Cox model, derived from penalized likelihood estimation. The estimates are consistent with previous results. However, our proposal improves the last results with more stability. In particular, note that all parameter estimates are finite. Table 4.5 gives for each variable an estimate, the hazard ratio and the sign of the association. In order to choose the best model we use different setting to build the model. We refer the reader to the previous Sections for the methodological details. We report the results on Stratified random effects Bayesian Cox Model Averaging based on our proposal on Bayesian Stratified Penalized likelihood estimation. In the table below, we run the model one for each λ. In particular, λ comes from our Bayesian procedures described before. In order to reduce the possibility of bias caused by the effect of starting values for λ, observations within an initial transient or burn-in period are discarded. We wait for our Markov Chain to reach a state of equilibrium, before

$Variable$	p	EV	$Model_1$	$Model_2$	$Model_3$
info activation	100	1.0977	1.1	1.1	1.1
β **info disconnection**	100	0.8904	0.89	0.9	0.88
technical problem	100	-0.5073	-0.51	-0.51	-0.51
contractual variation	5.7	-0.0071	.	.	.
Moovie package	8.9	0.0191	.	0.21	.
β_8 **special offers**	100	3.2182	3.2	3.2	3.2
special discount offers	100	2.9434	2.9	2.9	3.0
info administrative	100	1.5115	1.5	1.5	1.5
payment methods	4.2	0.0045	.	.	.
β_5 **promotion**	100	-0.9436	-0.94	-0.94	-0.96
Sport package	6.0	-0.0077	.	.	-0.13
payment with bancomat	100	0.7970	0.8	0.79	0.79
geographical area	100	0.4168	0.42	0.41	0.41
special card offers	100	3.4814	3.5	3.5	3.5
old customer	100	5.4121	5.4	5.4	5.4
β_i **...**	...	...	...	.	...
$Rental_1$	.	-1.4896	-1.5	-1.5	-1.5
$Rental_2$	.	0.0215	0.022	0.029	0.014
$Rental_3$	.	0.3732	0.37	0.37	0.37
$Rental_4$	.	1.2731	1.3	1.3	1.3
$Channel_1$	.	-1.0705	-1.1	-1.1	-1.1
$Channel_2$	.	-1.0963	-1.1	-1.1	-1.1
$Channel_3$	.	-0.7992	0.8	0.8	0.78
$Channel_4$	.	0.2743	-0.27	-0.28	-0.27
$Channel_5$	.	-1.2890	-1.3	-1.3	-1.3

Table 4.3: Fixed effects Stratified Cox Bayesian Model averaging

$Model$	$PosteriorProbability$	$nVar$
$Model_1$	0.752	13
$Model_2$	0.089	14
$Model_3$	0.06	14
$Model_4$	0.057	14
$Model_5$	0.042	14

Table 4.4: Stratified fixed effects Cox Bayesian Model averaging: the best 5 models

$Variable$	$Estimate$	HR	$Association$
info activation	1.2984	3.66	+
β info disconnection	0.7379	2.092	+
technical problem	-0.5144	0.598	-
contractual variation	-0.0848	0.919	NA
Moovie package	0.2071	1.230	NA
β_8 special offers	2.9506	19.118	+
special discount offers	2.8246	16.855	+
info administrative	1.5225	4.584	+
payment methods	0.1615	1.175	NA
β_5 promotion	-1.118	0.329	-
Sport package	-0.058	0.943	NA
payment with bancomat	0.7228	2.060	+
geographical area	0.3871	1.473	+
smart card offers	4.32	75.539	+
old customer	6.1058	91.194	+

Table 4.5: Stratified random effects Cox Model: results

we use the output form the chain as the basis of inference. Since the equilibrium distribution is independent of the initial state of the chain, one way to determine whether or not the chain has converged to its equilibrium distribution is to try and measure the chains dependence on its starting value. When this dependence is essentially zero, we say that the chain has converged. To choose our λ we run a number of chains in parallel, and from different starting points, and then essentially using an Analysis of Variance technique to assess whether or not each of the chains have the same distribution. Specifically, the method involves calculating the variance of the means of the chains,

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\lambda}_{.j} - \bar{\lambda}_{..})^2, \tag{4.107}$$

where λ_{ij} denotes the state of chain j at time i, and we run m chains. We call this the between chain variance. We then calculate the within chain variance, W, which is the mean of the variances of each the chains, i.e,

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \tag{4.108}$$

where $s_j^2 = \frac{1}{n-1}\sum_{i=1} n(\lambda_{ij} - \bar{\lambda}_{.j})^2$. Finally, we calculate the weighted average,

$$\hat{V} = \frac{n-1}{n}W + \frac{B}{m}, \tag{4.109}$$

and monitor the ratio:

$$R = \frac{\hat{V}}{W}, \tag{4.110}$$

until it reduces to the value 1. With this method is possible to examine the different chains and tries to measure the variation associated with the chains. There is variation associated with each chain individually (since these are stochastic simulations) but there is also some variation between the chains, because they were started in different states. The chains have converged when the variation between the chains is very small (since they will all be samples from the target distribution), so the ratio R will be close to 1. Here we show the best results for λ. As starting value we use the best λ that comes from the best Classical Penalized model and we run our algorithm for different iteration. First we use a very simple proposal distribution, as an Uniform distribution. We use 1000 observations as burn in period. For feature research we will compare the results with a new algorithm that uses the random walk metropolis algorithm to sample the posterior of a gamma distribution using

$Model$	λ	$Iterations$	$Variance$
P_1	2.7166	5000	0.9728
P_2	1.9297	50000	0.8631
P_3	1.9766	500000	0.7589

Table 4.6: Bayesian smoothing choice

uniform proposals. In Table 4.6 we report the results based on the previous values of smoothing parameter. We take also the mean for each estimated parameter in order to have an expected value for each variable. In order to choose the best penalized model in Table 4.7 we show the results, in Table 4.8, based on the log-Likelihood. We then plot in Figure 4.5 the hazard function that comes from the best Random effects stratified hazard models. As we can see in Figure 4.5, the hazard is

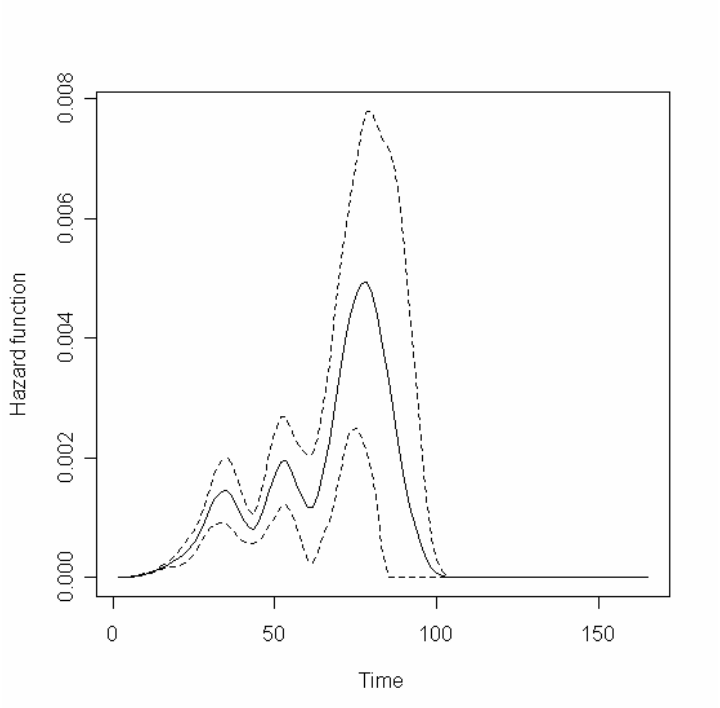

Figure 4.3: Random effect stratified hazard function

not a monotonic function. The distribution in the figure present three modes. This suggest further investigation to understand if in our population there are different groups of customers; this may suggest the usage of Mixture Models. Finally, from

$Variable$	p	PEV	P_1	P_2	P_3
info activation	100	1.3201	1.168	1.255	1.2702
β info disconnection	100	0.7708	0.784	0.7310	0.7365
technical problem	100	-0.5156	-0.496	-0.5106	-0.5112
contractual variation	100	-0.0803	-0.103	-0.0863	-0.0880
Moovie package	100	0.2042	0.209	0.2209	0.2168
β_8 special offers	100	2.9431	2.267	2.7830	2.9031
special discount offers	100	2.8189	2.140	2.6570	2.7798
info administrative	100	1.5254	1.532	1.5248	1.5274
payment methods	100	0.1560	0.128	0.1680	0.1731
β_5 promotion	100	-1.1077	-1.113	-1.1409	-1.1136
Sport package	100	-0.0580	-0.052	-0.0553	-0.0463
payment with bancomat	100	0.7331	0.728	0.7260	0.7368
geographical area	100	0.3912	0.384	0.3930	0.3903
smart card offers	100	4.3229	3.605	4.1464	4.2804
old customer	100	4.5181	4.097	5.1751	4.8789
β_i **...**	...	...	...	...	...

Table 4.7: Stratified random effects Bayesian Cox Model Averaging

$Model$	$PMLL$	$nVar$
PEV	-1969.24	15
P_1	-1985.9	15
P_2	-1983.55	15
P_3	-1974.6	15

Table 4.8: Stratified random effects Cox Model Averaging: the best 4 models

the models in Table 4.7 it is possible to derive the relative survival function that is close by the confidence interval, especially before 100 months of activation. The results are shown in Figure 4.6.

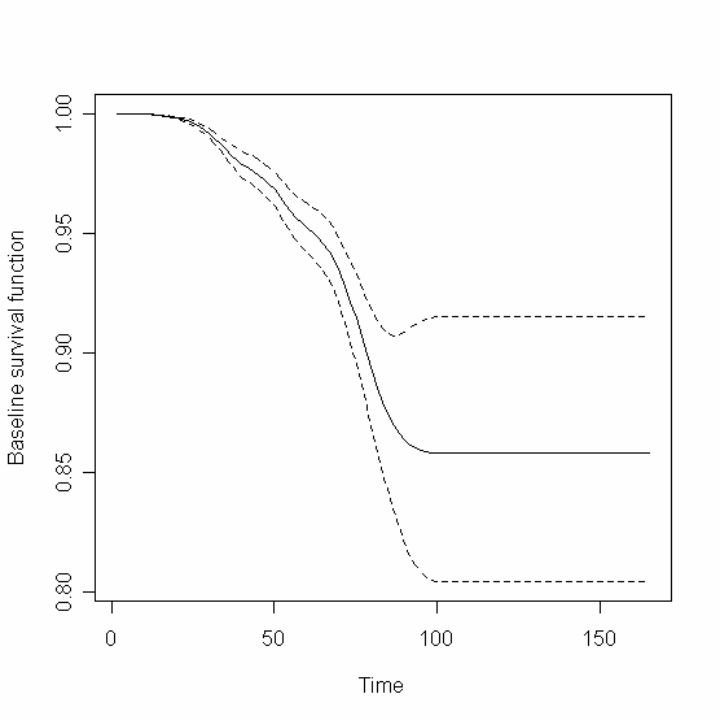

Figure 4.4: Random effect stratified hazard function

Chapter 5

Bayesian model assessment

In this chapter we propose a new method to classify Bayesian model selection results with a cluster approach. We remark that this is only an empirical proposal. The first part of this Chapter must be improved. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Second we propose a statistical and economical approach to choose the best model. We then apply the methodology to our proposed application for model selection.

5.1 Distance-based methods for model choice

We focus on a new methodological procedure to measure the quality of Bayesian model selection procedures. We start with a very simple remark. If we look at the results that comes from the Bayesian models averaged, in our case, two models often differ only for one or two variables (in terms of number of variables included).
We therefore point attention on some new methods to reduce model dimensionality. From the Bayesian model averaging approach, we can have available $2^p = m$ models, where p is the number of variables. But, some models are very similar. The similarity can be measured by a distance function. Suppose that we can derive a 0/1 matrix,Z, where the columns are the 2^p models and p are the number of rows. We put 1 in the matrix if the variable p_1 is present in Model M_1, otherwise we put 0. We report here a simple example. For example, if $x_{11} \in Z$ is 1 and $x_{1m} \in Z$ is 0, the variable 1 is present i model 1 and not in model m. For each column of Z we can observe the

presence-absence for each variable in each model.

$$\mathcal{Z} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pm} \end{pmatrix}$$

We would like to aggregate similar models in order to have a small set of different models. There are many ways in which a distance or a proximity may be measured starting from the Z matrix. In our case, a similarity coefficient indicates the strength of the relationship between two objects (models) given the presence or the absence of values of a set of p variables common to both.

The similarity between two objects i and j, will be some function of their observed values, i.e.

$$s_{ij} = f(x_i, x_j), \tag{5.1}$$

where $\acute{x_i} = [x_{i1}, ..., x_{ip}]$ and $\acute{x_j} = [x_{j1}, ..., x_{jp}]$ are the observed variable values for the objects. Many functions have been proposed depending partly on the type of variable concerned (quantitative, categorical, binary, ordinal) and partly on the type of object.

Similarity is usually regarded as a symmetric relationship requiring $s_{ij} = s_{ji}$. Asymmetric proximity measures are considered in Constantine and Gower, 1978. Most similarity coefficients are non negative and are scaled so as to have an upper limit of unity, although some are of a correlational nature so that:

$$-1 \leq s_{ij} \leq 1. \tag{5.2}$$

Associated with every similarity measure bounded by zero and unity is a dissimilarity (distance) $d_{ij} = 1 - s_{ij}$ which is symmetric and non negative. The degree of similarity between two objects increases with s_{ij} and decreases with d_{ij}. It is natural for an object to have maximal similarity with itself so that $s_{ij} = 1$ and $d_{ij} = 0$. In our case, we consider the Z matrix that is related to the presence or absence of some quality (variable).

The distance between two models, i and j, contained in the space of the models, M, where $dim(M) = m$, can be arranged in a 2×2 table as in Table 5.1. Such table contains the joint frequencies that count how many variables appears, respectively,

	$M_i = 0$	$M_i = 1$	$Total$
$M_j = 0$	a	b	a+b
$M_j = 1$	c	d	c+d
$Total$	a+c	b+d	p

Table 5.1: Frequency counts for two models

$Variable$	M_1	M_2	M_3	...	M_m
Var_1	1	0	0	...	...
Var_2	0	0	0	...	...
Var_3	0	0	0	...	...
Var_4	0	0	0	...	...
Var_5	1	1	0	...	...
Var_6	1	0	0	...	...
Var_7	0	0	0	...	...
Var_8	0	1	1	...	...
Var_9	1	1	0	...	...
Var_{10}	0	0	0	...	...
...	...	...	...	...	...
...	...	...	...	...	...
Var_p	...	...	...	...	...

Table 5.2: Models comparison: an example

in no model ($M_i = 0, M_j = 0$), one model only ($M_i = 1, M_j = 0$); ($M_i = 0, M_j = 1$) and both models ($M_i = 1, M_j = 1$). Many similarity coefficients have been proposed that combine the quantities a,b,c and d, see e.g. Sneath and Sokal (1973), Anderberg (1973), Clifford and Stephenson (1975), Cormack (1971) and Gower (1985). The two coefficients most commonly used in practice are the matching coefficient, $\frac{a+d}{p}$ and the Jaccard coefficient, $\frac{a}{a+b+c}$. Sneath et al. (1973) give a full discussion of similarity coefficients for use with binary data and argue that no rule can be made regarding the inclusion or otherwise of negative matches.

We remark that different similarity coefficients may have different values for the same set of data. Suppose for example that we consider the first 10 variables of our dataset

M_2/M_1	$M_1 = 1$	$M_1 = 0$	$Total$
$M_1 = 1$	2	1	3
$M_1 = 0$	2	5	7
$Total$	4	6	10

Table 5.3: Pairwise model comparison

$Index$	$Value$
$\frac{a+d}{p}$	0.70
$\frac{a}{a+b+c}$	0.40
$\frac{2a}{2a+b+c}$	0.57
$\frac{2(a+d)}{2(a+d)+b+c}$	0.82
$\frac{a}{a+2(b+c)}$	0.25

Table 5.4: Similarity coefficients values

and compare the first two model selected by the one step Bayesian model averaging approach. The corresponding 2×2 table of counts is presented in Table 5.3. In Table 5.4 we show the values taken by the various similarity coefficients. That the different coefficients take different values for the same pair of models would be relatively unimportant if the coefficients where jointly monotonic, in the sense that, if all the values for different pairs of models on one coefficient were ordered so that they formed a monotonic series (that is a series which either increases or decreases throughout its length), the corresponding values for other coefficients were similarity ordered. That this is not necessarily the case is most easily demonstrated by introducing data on the ten binary variables considered previously, for a further model. We consider M_3 and compare the first three models with a similarity coefficient. In Table 5.5 we present results for the simple matching coefficient and the Jaccard coefficient. Notice that the coefficients are not jointly monotonic. In order to obtain monotonicity it is possible to attach scores s_{ijk} of zero or one, to each variable, k, $k = 1, ..., p$, depending on whether the two models i and j have the same binary value on that variable: that is if both include or not include that variable. The scores for all variables are then simply averaged to obtain a similarity coefficient as $M_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{p}$. This can ensure monotonicity (non-invariant similarity).

Model	*Matching*	*Jaccard*
M(1,2)	0.70	0.40
M(1,3)	0.50	0.00
M(2,3)	0.80	0.33

Table 5.5: Multiple model comparison

If we like to have an immediate comparison among models, we use the complement of similarity measures called dissimilarity matrix. This matrix, that we call D stores a collection of proximities that are available for all pairs of m models. It is often represented by a $m \times m$ table as shown below:

$$\mathcal{D} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1m} \\ d_{21} & 0 & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & 0 \end{pmatrix}$$

where $d(i, j)$ is the measured difference or dissimilarity between model i and model j. Since $d(i, j) = d(j, i)$ and $d(i, i) = 0$ we have the matrix in D.
For binary data, Gower (1966) has shown that:

$$d_{ij} = \sqrt{2(1 - sij)}, \tag{5.3}$$

can function as an Euclidean distance provided the matrix od similarity coefficients in positive semi definite. Both the simple matching coefficient and the Jaccard coefficient meet this requirement. Although the Euclidean distance is the most widely used in a clustering context, other distance measures have been employed (the city block metric, the Canberra metric, the Angular separation, the Mahalanobis distance and so on). In our proposal we use two different way to reduce the number of possible models based on graphical displays of distance matrices and cluster analysis.

5.1.1 Multidimensional model scaling

The information about similarities or distances in a similarity or a distance matrix can be presented graphically, usually by deriving coordinate values for the models, which may then be plotted. Such plots might then be examinated for evidence of distinct groups of points, or they might be used in association with the results of

some grouping technique.

The most common approach to finding a coordinate representation of a set of distances or similarities is by using one of a collection of techniques known as multidimensional scaling. Such methods are described in detail in Everitt and Dunn (1991). In our proposal, a geometrical or spatial representation of the observed proximity matrix consists of a set of points $x_1, .x_2, ..., x_m$ in p dimensions, each point representing one of the models under investigation, and a measure of distance between pairs of points.

The objective of multidimensional scaling is to determine both the dimensionality needed to represent the information in the proximity matrix adequately, and the position of the points so that there is, in some sense, maximum correspondence between the observed proximities and the interpoint distances. In general terms this simply means that the larger the observed distance between two models the further apart should be the points representing them.

The required coordinates are found by minimizing some function which measures the discrepancy between the observed proximities and the fitted distances. Many such functions have been suggested, a number of which are described in Everitt and Dunn (1991). In all cases the results will consist of a set of coordinate values for each model. The hope is that the first few of these will provide an adequate representation of the observed proximities.

Multidimensional model scaling procedures are generally iterative optimization procedures. They start with some type of initial solution and then improve the solution in steps. When the solution cannot be improved, the procedure stops. In each iteration there are two phases: one phase derives a new configuration (coordinates of the objects in a low dimensional space) and the other phase is called the optimal scaling phase. The measurement level of the data determines what takes place during the optimal scaling phase.

We now propose two simple applications of Multidimensional scaling to model selection. We call this multidimensional model scaling (MDMS). We consider the proximity matrix that come from the Bayesian Stratified model averaging presented in Section 4.9 and we consider only the first five models selected. The process is as follows:

1. Calculate proximities as described in the section before;

2. Derive an initial configuration;

3. Perform an optimal scaling conditioned on the configuration, to try to improve the fit of the model to our data;

4. Test to see whether the fit has improved sufficiently to continue. If it has not, we exit, otherwise we continue to the next step;

5. Derive a new configuration, conditioned on the results of the optimal scaling. We move the points around in the space defined by the axes to try to improve the fit of the model to our data.

6. Return to step 3.

We finally produce a plot of coordinate values and we add labels to the plot to identify the points and draw axes such that one unit on the vertical and horizontal axes is the same distance. When axes are equated in this manner, distances are correctly presented in the plot. The plot shows the relative location of models. Because we

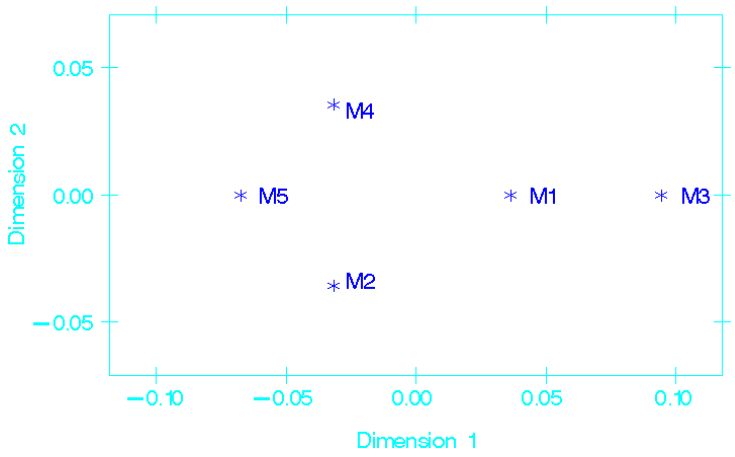

Figure 5.1: Dimensional scaling

use for the data the absolute value, the distances between the coordinates that we plot are on the same scale as the distances. Figure 5.1 presents the results of the procedure. We adopt a particular case of Table 5.2, where M_1,...,M_5 are selected from the Bayesian Stratified model averaging presented in Section 4.9. Based on Jacard coefficient, we plot the relative proximity matrix.

As we can see M_2, M_5 and M_4 are very near and differ from M_1 ,M_3. From an empirical point of view, this suggests some similarity between there models and some ideas on grouping models.

If we run our algorithm for a larger number of models, (we remember that from the Bayesian Model Averaging procedure it is possible to derive $m = 2^p$ models, where p is the number of covariates available) the plot is more helpful.

Our idea is to aggregate such similar models with clustering methods.

The classic clustering technique is called k-means. First, it is necessary to specify in advance how many clusters are being sought: this is the parameter k. Then k points are chosen at random as cluster centres. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metric. Next the centroid, or mean, of the instances in each cluster is calculated this is the "means" part. These centroids are taken to be new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain the same forever. This clustering method is simple and effective. It is easy to prove that choosing the cluster center to be the centroid minimizes the total squared distance from each of the cluster points to its center. Once the iteration has stabilized, each point is assigned to its nearest cluster center, so the overall effect is to minimize the total squared distance from all points to their cluster centers. But the minimum is a local one; there is no guarantee that it is the global minimum. The final clusters are quite sensitive to the initial cluster centers. Completely different arrangements can arise from small changes in the initial random choice. In fact, this is true of all practical clustering techniques: it is almost always infeasible to find globally optimal clusters. To increase the chance of finding a global minimum often the algorithm is run several times with different initial choices and choose the best final result the one with the smallest total squared distance.

The objective of model cluster analysis, is to cluster the model observations in groups, internally homogeneous (internal cohesion) and heterogeneous between them (external separation). In our proposal we start with classical cluster procedures, but we put particular attention at some optimization methods that can improve the results. In particular we discuss optimization criteria to choose the number of clusters

and, consequently, model clustering configuration.

5.1.2 Optimization methods for cluster analysis

Associated to each partition of the m models into the required number of groups, g there is an index $f(m, g)$ which is indicative of the quality of this particular clustering. From a partition of the data we consider the following three matrices:

$$\begin{aligned} T &= \frac{1}{m}\sum_{i=1}^{g}\sum_{j=1}^{m_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})^T \\ W &= \frac{1}{m-g}\sum_{i=1}^{g}\sum_{j=1}^{m_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})^T \\ B &= \sum_{i=1}^{g} m_i(x_i-\bar{x})(x_i-\bar{x})^T \end{aligned} \tag{5.4}$$

These $p \times p$ matrices, where p is the number of variables, represent respectively total dispersion, within group dispersion and between group dispersion, and satisfy the following equation:

$$T = W + B. \tag{5.5}$$

For $p = 1$ this equation represents a relationship between scalars; simply the division of the total sum of squares for a variable into the within and between groups of sum of squares, familiar from a one way analysis of variance. In this case a natural criterion for grouping would be to choose the partition corresponding to the minimum value of the within group sum of squares, or equivalently, the maximum value of the between group term. For $p > 1$ the derivation of clustering criteria from the previous equation is not so clear cut, and several alternatives have been suggested:

- Minimization of the W trace, see e.g. Singleton and Kautz (1965), Forgey (1965), Jancey (1966), MacQueen (1967) and Ball and Hall (1967).
- Minimization of the determinant of W, see e.g. Krzanowsky (1988), Friedman and Rubin (1967), Marriott (1971,1982).
- Maximization of the trace of BW^{-1}, see e.g. Wallace and Boulton (1968), Rao (1952), Spath (1985).

Once a suitable numerical clustering criterion has been selected, consideration needs to be given on how to choose the group partition g of the data which leads to its

optimization.

In order to select the number of groups a number of formal techniques are available. Beale (1969) for example, gives an "F test", which may be used to test whether a sub division into g_2 clusters is significantly better than a subdivision into some smaller number of clusters, g_1. The test statistic is defined as follows:

$$F(g_1, g_2) = \frac{R_{g1} - R_{g2}}{R_{g2}} \left[\left\{ \frac{m - g_1}{m - g_2} \right\} \left(\frac{g_2}{g_1} \right)^{2/p} - 1 \right]^{-1} \tag{5.6}$$

where $R_g = (m - g)S_g^2$ and S_g^2 is the mean square deviation from cluster centers in the sample. The statistic is compared with the F statistic, F with $p(g_2 - g_1)$ and $p(m - g_2)$ d.f. Experience with this procedure suggests that it will only be useful when the clusters are fairly well separated and approximately spherical in shape.

A method suggested by Calinsky and Harabasz (1974) is to take the value of g which corresponds to the maximum value of C, where C is given by:

$$C = \frac{trace(B)}{g - 1} \left(\frac{trace(W)}{m - g} \right)^{-1} \tag{5.7}$$

This criterion performs reasonably well in the study of indicators of number of groups reported in Milligan and Cooper (1985).

Marriot (1971) suggests as another possible procedure for assessing the number of groups, to take the value of g for which $g^2 det(W)$is a minimum.

We remark that this research is only based on an empirical approach. We will improve this field of research from a theoretical point of view.

5.2 Predictive performance

Now, in order to choose the best model, we focus on predictive performance. One of the main arguments for using the BMA is based on its ability to improve our predictions, as measured by the out-of-sample prediction error.

BMA was successfully applied to various statistical methods of data analysis: univariate linear regression (Raftery et al., 1997; George and McCulloch, 1993; Geweke, 1996), multivariate analysis (Brown and Vannucci, 1998; Noble, 2000), survival analysis (Volynsky et al., 1997), generalized linear models (Raftery, 1995; Clyde, 1999), graphical models (Madigan and Raftery, 1994), wavelet estimation (Clyde and George, 1999), regression trees (Chipman et al., 1998), and nonparametric regression (Smith and Kohn, 1996; Shively et al., 1999).

In Hoeting et al. (1999) we see several examples of BMA applications, each equipped with a convincing out-of-sample validation in terms of its predictive performance. The cross-validation was performed by splitting each data set into two parts, training set, D^T and prediction set, D^P. The training set is used for model selection and the second set for prediction. Using several models turned out better than using a single model in most of the cases. Two measures of predictive ability were used, the coverage for a 90% predictive interval, measured by proportion of observation of the second set falling within the 90% of the corresponding posterior prediction interval (see Hoeting et al., 1999). The second measure is the logarithmic scoring rule of Good (1952). Specifically we measure the predictive ability of a single model M as:

$$-\sum_{\Delta \in D^P} log\left\{pr(\Delta|M, D^T)\right\}, \tag{5.8}$$

And compare it with predictive ability of BMA as measured by:

$$-\sum_{\Delta \in D^P}\left[log\left\{\sum_{k=1}^{k} pr(\Delta|M_k, D^T)pr(M_k|D^T)\right\}\right] \tag{5.9}$$

The smaller the predictive log score for a given model or model average, the better the predictive performance. An intuitive explanation of such a good performance of BMA was given in George (1999), who noted that BMA based prediction can be viewed as an application of averaging of several approximately unbiased estimates with weights adaptively accounting for their varying variance, hence better prediction. A more analytical argument was used in Madigan and Raftery (1994), which follows from the non-negativity of the Kullback-Leibler information divergence.

$$-E\left[log\left\{\sum_{k=1}^{k} pr(\Delta|M_k, D)pr(M_k|D)\right\}\right] \leq -E\left[log\left\{pr(\Delta|M_j, D)\right\}\right], j = 1, ..., k \tag{5.10}$$

To assess the performance of BMA in the case of Bayesian survival models. I carried out a simulation study. The idea of simulation was to use some "true" model to generate the data and then compare the out of sample prediction for the averaged model to that of various candidate models (full model, top selected model, and true model). Of course, the usefulness of the results is limited because the simulation is based on the assumption that a "true" model exists, which may not be the case in a real life situation. The simulation procedure was as follows.

Training and prediction data sets were simulated under the same generation scheme,

D^B and D^T. The simulation was carried out for fixed and random X designs. For the fixed design we used same (generated) values of explanatory variables for both training and prediction data sets. We assume that the full model is not the true model. First, we apply each model selection method to the first part of the data, D^B. The corresponding coefficient estimates define a predictive density for each person in the second part of the data D^P. Then a log score (see Good, 1952) for any given model M_k is based on the observed ordinate of predictive density for the subjects in D^T:

$$\sum_{d \in D^T} logpr(d|M_k, D^B). \tag{5.11}$$

Similarly the predictive log score for BMA is:

$$\sum_{d \in D^T} log \left\{ \sum_{m \in M} pr(d|M, D^B) pr(M|D^B) \right\}, \tag{5.12}$$

where M is the set of BMA selected models.

However, the Cox model does not directly provide a predictive density. Rather it provides an estimated predictive CDF which is a step function (Breslow, 1975) and therefore does not lead to differentiation into a density. In the spirit of Cox partial likelihood we have designed an alternative to the predictive density:

$$pr(d|M_k, D^B) = \left(\frac{exp(x_i \hat{\theta}_k)}{\sum_{l \in R_i} exp(x_l^T \hat{\theta}_k)} \right)^{\delta_i}. \tag{5.13}$$

By substituting this in the two equations before, we obtain an analogue to a log score called the partial predictive score (PPS). Using the PPS we can compare BMA to any single model selected. The partial predictive score is greater for the method which gives higher probability to the events that occur in the last test set.

We also compare methods based on their predictive discrimination, namely how well they sort the subjects in the test set into discrete risk categories (high, medium, low risk). We assess predictive discrimination of a single model as follows:

- Fit the model to build data to get estimated coefficients $\hat{\theta}$.
- Calculate risk scores $(x_i^T, \hat{\theta})$ for each subject in the build data.
- Define low, medium and high risk groups for the model by empirical percentiles of risk scores (in the simplest case, 33rd and 66rd)
- Calculate risk scores for the test data and assign each subject to a risk group.

- Extract the subjects who are assessed as being in a higher risk group by one method than by another, and tabulate what happened to those subjects over the study period.

To assess predictive discrimination for BMA, we must take account of all of the averaged models. We replace the first step above with:

- Fit each model $M_1, ..., M_k$ in A, where A is a set where $\frac{max_l(pr(M_l|D))}{pr(M_k|d)} < C$ (Raftery, 1997, show that $C = 20$ provides a good approximation to averaging over the entire model space), to get estimated coefficients, $\hat{\theta}_k$.

- Calculate risk scores $(x_i^T, \hat{\theta}_k)$ under each model in A for each person in the build data. A score risk of person under BMA is the weighted average of these: $\sum_{k=1}^{K}(x_i^T, \hat{\theta}_k)pr(M_k|D^B)$.

A method is better if it consistently assigns higher risks to the people who actually had the event at risk.

In order to apply what presented, we split the data randomly, the training data set and the validation data set. In order to measure the predictive discrimination we use a confusion matrix. The confusion matrix is used as an indication of the properties of a classification (discriminant) rule. It contains the number of elements that have been correctly or incorrectly classified for each class. On its main diagonal we can see the number of observations that have been correctly classified for each class while the off-diagonal elements indicate the number of observations that have been incorrectly classified. If it is (explicitly or implicitly) assumed that each incorrect classification has the same cost, the proportion of incorrect classifications over the total number of classifications is called rate of error, or misclassification error, and it is the quantity which must be minimised. Of course the assumption of equal costs can be replaced by weighting errors with their relative costs. If there are different costs for different errors a model with a lower general level of accuracy is preferable to one that has greater accuracy but also much higher costs.

We have calculated predictive performance for the one-step Bayesian model averaging procedure presented in this dissertation. Partial predictive scores for the top models (as determined from the build set) in Table 3.6 are $PPS = -634.6$ for the model with the highest posterior probability, $PPS = -423.6$ for the model averaging. Model averaging performs better than the single model methods. For the

O/P	$P = 0$	$P = 1$	$Total$
$O = 0$	a	b	a+b
$O = 1$	c	d	c+d
$Total$	a+c	b+d	a+b+c+d

Table 5.6: Theoretical Confusion Matrix

top models in Table 3.5, there is $PPS = -333.2$ for the model with the highest posterior probability, $PPS = -278.6$ for the model averaging. We report here an example of confusion matrix for the best predictive model in Table 3.6. We consider a sample of 1000 customers. Table below classifies the observations of a validation dataset in four possible categories: the observations predicted as events and effectively such (with absolute frequency equal to a); the observations predicted as events and effectively non events (with frequency equal to c); the observations predicted as non events and effectively events (with frequency equal to b); the observations predicted as non events and effectively such (with frequency equal to d). We obtain the following results: $a = 645$, $b = 45$, $c = 134$, $d = 176$. The frequency in b and c represents two different type of error. It is possible also to derive the proportion of non events predicted as events (type II error, false positives) as $\frac{c}{c+d}$ or proportions of events predicted as non events (type I error, false negatives) as $\frac{b}{a+b}$. Also we can calculate the sensitivity as $\frac{a}{a+b}$ (proportion of events, predicted as such) and the specificity as $\frac{d}{c+d}$ (proportion of non events, predicted as such).

5.3 Lift of a model

In a prediction step the goodness of the model will be evaluated in terms of predictive accuracy in a cross-validation exercise. We have split the dataset in the two usual subsets: training and test. Both have been proportionally sampled, with respect to the status variable. All sampled data contain information on all finally chosen explanatory variables (about twenty). In order to evaluate the predictive performance of the model, and compare it with classification trees (routinely used by the company), we have focused our attention to the 3 months ahead prediction. We have implemented a procedure based on the estimated survival probabilities, aimed at building the confusion matrix (see e.g. Giudici 2003) and, correspondingly, the

percentage of captured true churners of the model. We remark that this is indeed not a fair comparison as survival models predict more than a point; however typically ask for this type of model benchmarking.
In correspondence of each estimated probability decile, the percentage of true churners captured is : 0.0504 in the first decile, 0.0345 in the second decile. In general, while in the first decile (that is, among the customers with the highest estimated churn probability) 0.05 of the clients are effective churners, the same percentage lowers down in subsequent deciles, thus giving an overall picture of good performance of the model. Indeed the lift of the model, as measured by the ratio between the captured true responses, between the model and a random allocation does not turn out to be substantially better with respect to what obtained with tree models. However, we remark that, differently from what occurred with the latter models, the customers with the highest estimated churn rate are now not necessarily those whose contract is close to the deadline. This is the most beneficial advantage of the survival analysis approach, which, in turn, leads to substantial gains in campaign costs.

5.4 ROC Curve

In this section we compare the predictive power of different models, namely Bayesian Model Averaging Stratified Models and Bayesian Penalized models.
We employ the same data set as training and validation in order to better compare the results. We cannot use the Partial Predictive Scores, because methodologically it is not so easy to derive it for the Penalized Likelihood estimation. We compare model performances on the basis of the confusion matrix described before. In order to better compare this two models, we calculate also new measures, based on the balanced error rate, BER $BER = (\frac{c}{a+c} + \frac{b}{b+d})/2$, $2 \times a - b$ that is a measure used in the filtering evaluation and $F_1 = \frac{2 \times a}{(2 \times a)+b+c}$, the F-measure of Van Rijsbergen with the trade-off parameter set to 1 (Van Rijsbergen, 1972).
In our case it turns out that the Bayesian Penalized model is the better model to estimate Customer Lifetime Value. We build also a ROC curve for model comparison. ROC graphs are another way besides confusion matrices to examine the performance of classifiers (Swets, 1988). A ROC graph is a plot with the false pos-

itive rate on the X axis and the true positive rate on the Y axis. The point (0,1) is the perfect classifier: it classifies all positive cases and negative cases correctly. It is (0,1) because the false positive rate is 0 (none), and the true positive rate is 1 (all). The point (0,0) represents a classifier that predicts all cases to be negative, while the point (1,1) corresponds to a classifier that predicts every case to be positive. Point (1,0) is the classifier that is incorrect for all classifications. In many cases, a classifier has a parameter that can be adjusted to increase TP at the cost of an increased FP or decrease FP at the cost of a decrease in TP. Each parameter setting provides a (FP, TP) pair and a series of such pairs can be used to plot an ROC curve. A non-parametric classifier is represented by a single ROC point, corresponding to its (FP,TP) pair. Figure 5.5 shows an example of an ROC graph with

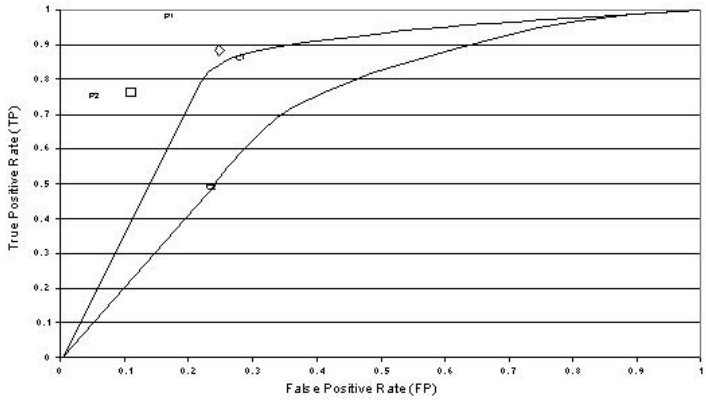

Figure 5.2: ROC curve comparison

two ROC curves labelled C1 (based on Bayesian Model Averaging Stratified Models) C2 (based on Bayesian Penalized model), and two ROC points labelled P1 and P2. Non-parametric algorithms produce a single ROC point for a particular data set. This graph can be improve. In fact, a ROC curve or point is independent of class distribution or error costs (Provost et al., 1998) but encapsulates all information contained in the confusion matrix, since FN is the complement of TP and TN is the complement of FP (Swets, 1988). The ideal point on the ROC curve would be (0,100), that is all positive examples are classified correctly and no negative examples are misclassified as positive.

We report here the properties for the ROC Curve: the slope is non-increasing, each point on ROC represents different trade-off (cost ratio) between false positives and

false negatives, the slope of line tangent to curve defines the cost ratio, the ROC Area represents performance averaged over all possible cost ratios. In order to compare two or more models, if two ROC curves do not intersect, one method dominates the other and if two ROC curves intersect, one method is better for some cost ratios, and other method is better for other cost ratios. It has been suggested that the area beneath an ROC curve can be used as a measure of accuracy in many applications (Swets, 1988). Provost et al. (1998) argue that using classification accuracy to compare classifiers is not adequate unless cost and class distributions are completely unknown and a single classifier must be chosen to handle any situation. They propose a method of evaluating classifiers using a ROC graph and imprecise cost and class distribution information.

Another way of comparing ROC points is by using an equation that equates accuracy with the Euclidian distance from the perfect classifier, point (0,1) on the graph. We include a weight factor that allows us to define relative misclassification costs, if such information is available. We have derived the ROC curve from the corresponding confusion matrix (Kohavi et al., 1998), which contains information about actual and predicted classifications done by a classification system. We obtain the following results: $a = 445$, $b = 45$, $c = 106$, $d = 404$ for the Bayesian Penalized model and $a = 497$, $b = 87$, $c = 108$, $d = 308$ for the Bayesian Model Averaging Stratified Models. The confusion matrix based on the classical Cox model give low results in model prediction.

A natural alternative to an error rate for model comparison is a misclassification cost. Instead of designing a classifier to minimize error rates, the goal would be to minimize misclassification costs. A misclassification cost is simply a number that is assigned as a penalty for making a particular type of a mistake. For example, in the two-class situation, a cost of one might be assigned to a false positive error, and a cost of two to a false negative error. An average cost of misclassification can be obtained by weighing each of the costs by the respective error rate. Computationally this means that errors are converted into costs by multiplying an error by its misclassification cost.

Any confusion matrix has n^2 entries, where n is the number of classes. On the diagonal lie the correct classifications with the off-diagonal entries containing the various cross-classification errors. If we assign a cost to each type of error or misclassifica-

	$Predictedpositive$	$Predictednegative$	$Total$
$Positiveexamples$	TP	FN	Pos
$Negativeexamples$	FP	TN	Neg
$Total$	PPos	PNeg	N

Table 5.7: Economic confusion matrix

tion, the total cost of misclassification is most directly computed as the sum of the costs for each error. If all misclassifications are assigned a cost of l then the total cost is given by the number of errors and the average cost per decision is the error rate. By raising or lowering the cost of a misclassification, we are biasing decisions in different directions, as if there were more or fewer cases in a given class. Formally, for any confusion matrix, if E_{ij} is the number of errors entered in the confusion matrix and C_{ij} is the cost for that type misclassification, the total cost of misclassification is given by the equation below:

$$Cost = \sum_{i=1} \sum_{j=1} E_{ij} \times C_{ij}. \tag{5.14}$$

We have so far considered the costs of misclassifications, but not the potential for expected gains arising from correct classification. In risk analysis or decision analysis, both costs (or losses) and benefits (gains) are used to evaluate the performance of a classifier. A rational objective of the classifier is to maximize gains. The expected gain or loss is the difference between the gains for correct classifications and losses for incorrect classifications. In economic analysis utility theory is employed, which allows modification of risks by a function. The nature of this function is part of the specification of the problem and is defined before the classification problem is derived. In all these cases decisions are based on modified error rates so as to measure classifier performance in units typical for the problem domain, and also provide means for making more correct decisions.

Our proposal for model choice starts with a simple matrix that give us a first measure of assessment. We introduce some terminology. True Positive (TP) represents the correct classification, true negative (TN), the correct rejection, the false positive (FP) represents a false alarm (First type of error) and false negative (FN) is a misclassification error (Second type of error). In particular we observe that $positive/negative$ refers to prediction and $true/false$ refers to correctness.

The true positive rate, $tpr = TP/TP + FN$ is the fraction of positives correctly predicted and the false positive rate, $fpr = FP/neg = FP/FP + TN$ is a fraction of negatives incorrectly predicted $= 1 - TN/FP + TN$, where $TN/FP + TN$ is the true negative rate. We can define a first measure of accuracy for model selection in the following equation:

$$Accuracy = pos \times tpr + neg \times (1 - fpr), \tag{5.15}$$

which is a weighted average of true positive and true negative rates. As described before, we can plot the FP rate and the TP rate to obtain a ROC curve. As we

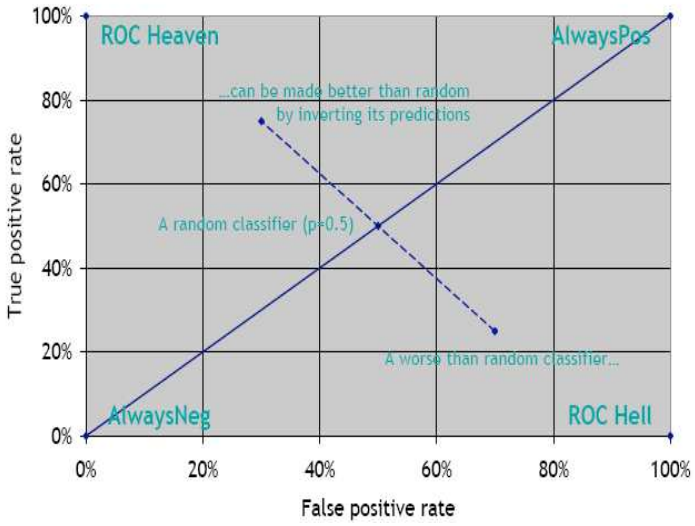

Figure 5.3: ROC curve comparison

can observe from Figure 5.5 we can plot a set of models (in our case 2 models) on a convex hull. In general, the classifiers on the convex hull achieve the best accuracy for some class distributions and classifiers below the convex hull are always sub-optimal.

Any performance on a line segment connecting two ROC points can be archived by randomly choosing between them: the ascending default performance is just a special case (absence of model). The classifiers on the ROC convex hull can be combined to form the ROCCH-hybrid (see e.g. Provost and Fawcett, 2001): there are the ordered sequence of classifiers that can be turned into a ranker. In particular, we can build an Iso accuracy line which connects ROC points with the same accuracy:

$$pos \times tpr + neg \times (1 - fpr) = a$$

$$tpr = \frac{a - neg}{pos} + \frac{neg}{pos} fpr \tag{5.16}$$

We can obtain a set of parallel ascending lines with slope neg/pos. The higher lines are better and on the descending diagonal, $tpr = a$. Each line segment on the convex hull is an iso accuracy line for a particular class distribution: under that distribution the two classifiers on the end points achieve the same accuracy, for distributions skewed towards negatives (steeper slope) the left one is better and finally, for distributions skewed towards positives (flatter slope) the right one is better. Each classifiers on convex hull is optimal for a specific range of class distributions. Now we want to incorporate costs and profits. The iso-accuracy and iso-error lines are the same. The error is a linear combination of $error = pos \times (1 - tpr) + neg \times fpr$ and the slope of iso-error line is neg/pos.

For each confusion matrix it is possible to derive easily the accuracy and the precision in prediction. In particular, the accuracy is a weighted average of true positive - negative rates, $accuracy = pos \times tpr + neg \times (1 - fpr) = \frac{tpr+c(1-fpr)}{c+1}$, where c is the skew ratio which indicates relative importance of negatives over positives and in the most simple case, without cost, $c = neg/pos$. The precision or confidence index is $precision = \frac{pos \times tpr}{pos \times tpr + neg \times fpr} = \frac{tpr}{tpr + c \times fpr}$. From this relation we can derive two variants, the relative precision, $(precision - pos)$ and the lift, $lift = precision/pos$. We derive also some test based on the F-measure. In particular, the F-measure is harmonica average of precision and some quantity derived from the confusion matrix. In ROC notation, $F = \frac{2 \times tpr}{tpr + c \times fpr + 1}$, equivalent but simpler $G = \frac{tpr}{c \times fpr + 1}$. We report here a summary for the linear metrics. We first introduce a Misclassification cost in order to obtain a total cost for a model performance.

$$cost = pos \times (1 - tpr) \times C(-|+) + neg \times fpr \times C(+|-), \tag{5.17}$$

The slope of iso-cost line is $neg \times C(+|-)/pos \times C(-|+)$. Now in order to incorporate the correct classification profits, we have:

$$\begin{aligned} cost = {} & pos \times (1 - tpr) \times C(-|+) + neg \times fpr \times C(+|-) + \\ & pos \times tpr \times C(+|+) + neg \times (1 - fpr) \times C(-|-), \end{aligned} \tag{5.18}$$

with a slope of iso-line $neg \times [C(+|-) - C(-|-)]/pos \times [C(-|+) - C(+|+)]$. But a ROC curve implicitly conveys information about performance across all possible combinations of misclassification costs and class distributions. We use the term

Metric	Formula	Skew-insensitive version	Isometric slope
Accuracy	$\frac{tpr + c(1 - fpr)}{c + 1}$	$\frac{(tpr + 1 - fpr)}{2}$	c
WRAcc*	$\frac{4c}{(c+1)^2}(tpr - fpr)$	$tpr - fpr$	1
Precision*	$\frac{tpr}{tpr + c \cdot fpr}$	$\frac{tpr}{tpr + fpr}$	$\frac{tpr}{fpr}$
Lift*	$\frac{c+1}{2}\frac{tpr}{tpr + c \cdot fpr}$	$\frac{tpr}{tpr + fpr}$	
Relative precision*	$\frac{2c}{c+1}\frac{(tpr - fpr)}{tpr + c \cdot fpr}$	$\frac{tpr - fpr}{tpr + fpr}$	
F-measure	$\frac{2tpr}{tpr + c \cdot fpr + 1}$	$\frac{2tpr}{tpr + fpr + 1}$	$\frac{tpr}{fpr + 1/c}$
G-measure	$\frac{tpr}{c \cdot fpr + 1}$	$\frac{tpr}{fpr + 1}$	

Figure 5.4: Linear metrics: summary

operating point to refer to a specific combination of misclassification costs and class distributions.

5.5 Model comparison based on ROC curves

We consider Figure 5.8. Although it is easy to see which curve is better it is much harder to determine by how much. This information can be extracted as it is implicit in the graph, but our alternative representation makes it explicit.

In general, one point in ROC diagram dominates another if it is above and to the left, i.e. has a higher true positive rate (TP) and lower false positive rate (FP). If a point A dominates point B, A will have a lower expected cost than B for all possible cost ratios and class distributions. One set of points A is dominated by another B when each point in A is dominated by the same point B when and no point in B is dominated by a point in A. The normal assumption in ROC analysis is that this points are samples of a continuous curve and therefore normal curve fitting techniques can be used. Alternatively a non-parametric approach is to use a piecewise linear function joining adjacent points by straight lines. Dominance is then defined for all points on the curve. The dashed line in Figure 5.9 is a typical ROC convex hull. The slope of a segment of the convex hull connecting the two vertices

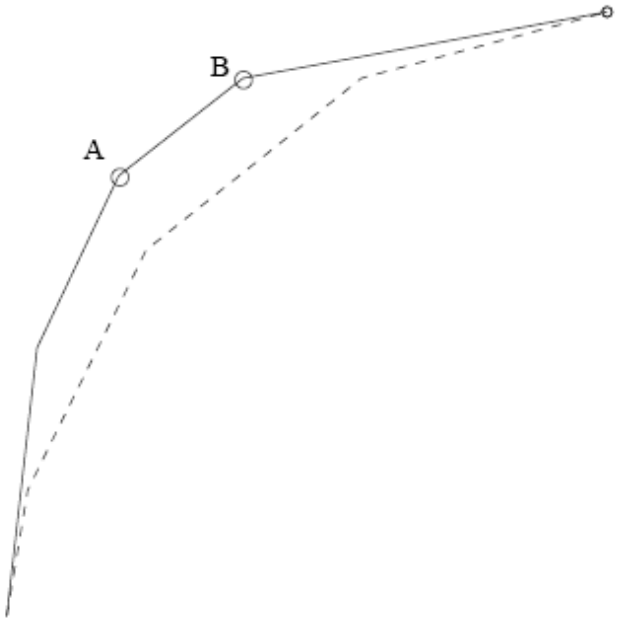

Figure 5.5: Roc choice

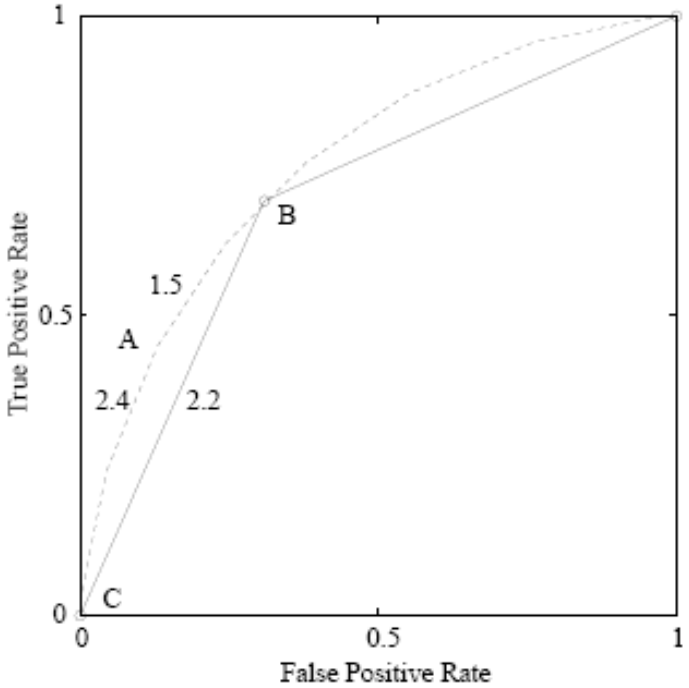

Figure 5.6: Roc convex hull

(FP_1, TP_1) and (FP_2, TP_2) is given by the following equation $\frac{TP_1 - TP_2}{FP_1 - FP_2} = \frac{p(-)C(+|-)}{p(+)C(-|+)}$, where $p(a)$ is the probability of a given example being in class a and $C(a|b)$ is the cost incurred if an example in class b is misclassified as being in class a. The previous equation defines the gradient of an iso-performance line. Classifiers sharing a line have the same expected cost for the ratio of priors and misclassification costs given by the gradient. Even a single classifier can form a ROC curve. The solid line in Figure 5.9 is produced by simply combining classifier B with the trivial classifiers: point $(0, 0)$ represents classifying all examples as negative; point $(1, 1)$ represents classifying all examples as positive. The slopes of the lines connecting classifier B to $(0, 0)$ and to $(1, 1)$ define the range of the ratio of priors and misclassification costs for which classifier B is potentially useful, its operating range. For probability cost ratios outside this range classifier B will be outperformed by a trivial classifier. As with the single classifier, the operating range of any vertex on a ROC convex hull is defined by the slopes of the two line segments connected to it. One of the questions posed in the introduction is how to determine the difference in performance of two ROC curves. For instance in Figure 5.9 the dashed curve is certainly better than the solid one. To measure how much better one might be tempted to take euclidean distance normal to the lower curve. But this would be wrong on two counts. Firstly, the difference in expected cost is the weighted Manhattan distance between two classifiers, given by the following equation, not the euclidean distance:

$$E[C_1] - E[C_2] = (TP_1 - TP_2)p(+)C(-|+) + (FP_1 - FP_2)p(-)C(+|-). \quad (5.19)$$

We call $p(+)C(-|+) = \omega_+$ and $p(-)C(+|-) = \omega_-$. Secondly the performance difference should be measured between the appropriate classifiers on each ROC curve. When using the convex hull these are the best classifiers for the particular cost and class frequency defined by the weights ω_+ and ω_- in the previous equation. In Figure 5.9 for a probability cost ratio of say 2.1 the classifier marked A on the dashed curve should be compared to the one marked B on the solid curve. But if the ratio was 2.3 it should be compared to the trivial classifier marked C on the dashed curve at the origin.

To directly compare the performance of two classifiers we transform a ROC curve into a cost curve. Figure 5.10 shows the cost curves corresponding to the ROC curves in Figure 5.9. The x axis in a cost curve is the probability cost function for positive

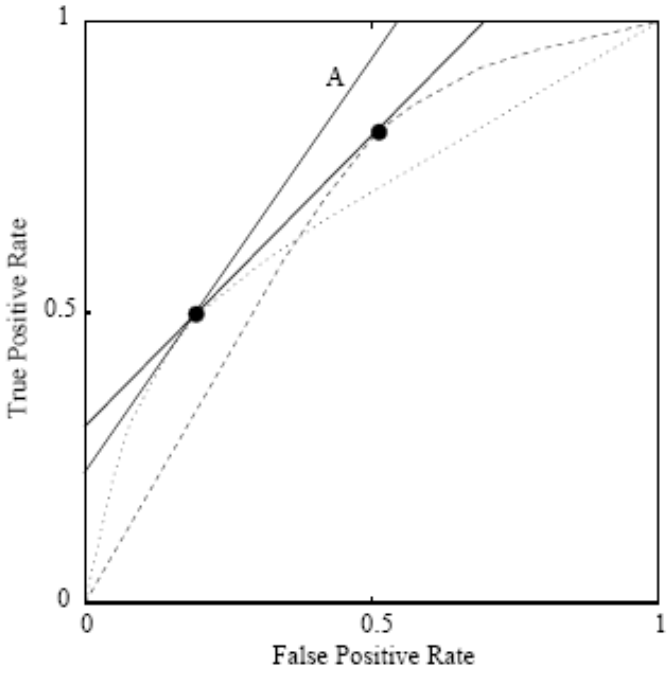

Figure 5.7: ROC space

examples, $PCF(+) = \omega_+/(\omega_+ + \omega_-)$ where ω_+ and ω_- are the weights in equation 5.19. This is simply $p(+)$, the probability of a positive example, when the costs are equal. The y axis is expected cost normalized with respect to the cost incurred when every example is incorrectly classified. The dashed and solid cost curves in Figure 5.10 correspond to the dashed and solid ROC curve in Figure 5.9. The horizontal line atop the solid cost curve corresponds to the classifier marked B. The end points of the line indicate the classifiers operating range ($0.3 \leq PCF(+) \leq 0.7$), where it outperforms the trivial classifiers. It is horizontal because $FP = 1 - TP$ for this classifier (see below). At the limit of its operating range this classifier cost curve joins the cost curve for the majority classifier. Each line segment in the dashed cost curve corresponds to one of the points (vertices) defining the dashed ROC curve. The distance between cost curves for two classifiers directly indicates the performance difference between them. The dashed classifier outperforms the solid one has a lower or equal expected cost for all values of $PCF(+)$. The maximum difference is about 20% (0.25 compared to 0.3), which occurs when $PCF(+)$ is about 0.3 (or 0.7). Their performance difference is negligible when $PCF(+)$ is near 0.5, less than

0.2 or greater than 0.8. It is certainly possible to get all this information from the ROC curves, but it is not trivial. The gradients of lines incident to a point must be brought into contact with each convex hull to determine which points must be compared. To find the actual costs the weighted Manhattan distance between them must be calculated. All this information is explicit in the alternative representation. The second question posed in the introduction was for what range of cost and class distribution is one classifier better than another. Suppose we have the two hulls in ROC space, the dotted and dashed curves of Figure 5.10. The solid lines indicate iso performance lines. The line designated A touches the convex hull indicated by the dotted curve. A line with the same slope touching the other hull would be lower and to the right and therefore of higher expected cost. If we roll this line around the hulls until it touches both of them we find points on each hull of equal expected cost, for a particular cost or class frequency. Continuing to roll the line shows that the hull indicated by the dashed line becomes the better classifier. It is noteworthy that the crossover point of the two hulls says little about where one curve outperforms the other. It only denotes where both curves have a classification performance that is the same but suboptimal for any costs or class frequencies. Figure 5.11 shows the cost graph that is the dual of the ROC graph of Figure 5.10. Figure 5.11 shows the cost graph that is the dual of the ROC graph of Figure 5.10. It can be immediately be seen that the dotted line has a lower expected cost and therefore outperforms the dashed line to the left of the crossover point and vice versa. This crossover point when converted to ROC space becomes the line touching both hulls shown in Figure 5.10.

To construct the alternative representation we use the normalised expected cost. The expected cost of a classifier is given by the following equation:

$$E[C] = (1 - TP)p(+)C(-|+) + FPp(-)C(+|-). \tag{5.20}$$

The worst possible classifier is one that labels all instances incorrectly so $TP = 0$ and $FP = 1$ and its expected cost is given by this equation:

$$E[C] = p(+)C(-|+) + p(-)C(+|-). \tag{5.21}$$

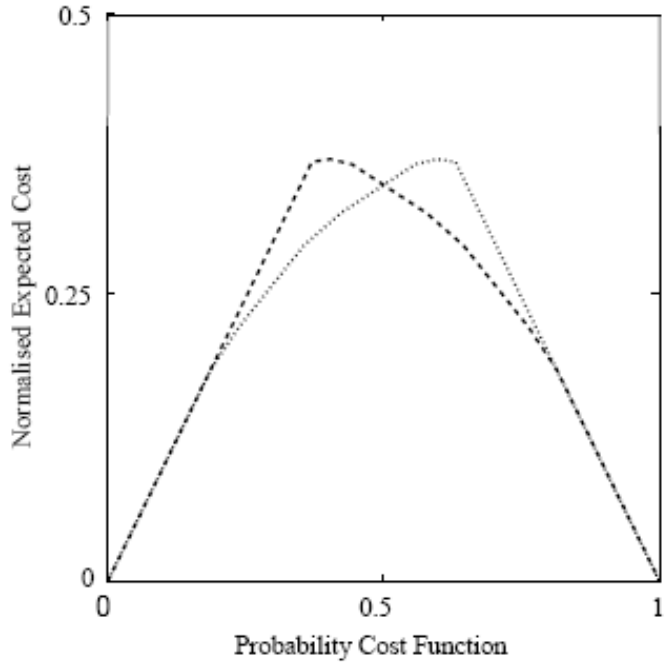

Figure 5.8: Cost space

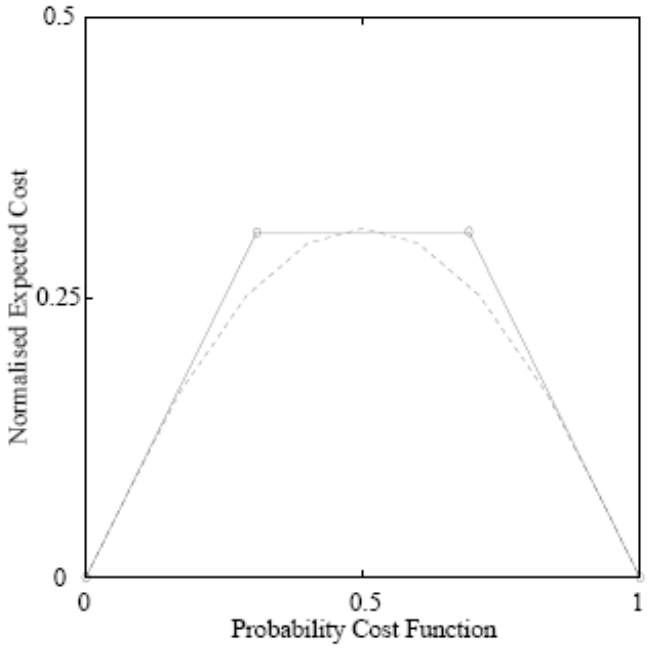

Figure 5.9: Cost space: specification

The normalised expected cost is then produced by dividing the right hand side of equation 5.20 by that of equation 5.21 given equation 5.22:

$$NE[C] = \frac{(1 - TP)p(+)C(-|+) + FPp(-)C(+|-)}{p(+)C(-|+) + p(-)C(+|-)}. \tag{5.22}$$

The replacing the normalised probability cost terms with the probability cost function $PCF(a)$ as in equation 5.23 results in equation 5.24:

$$PCF(a) = \frac{p(a)C(a|\bar{a})}{p(+)C(-|+) + p(-)C(+|-)}, \tag{5.23}$$

$$NE[C] = (1 - TP) \times PCF(+) + FP \times PCF(-). \tag{5.24}$$

Because $PCF(+) + PCF(-) = 1$, we can rewrite the last equation to produce equation 5.25 which is the straight line representing the classifier.

$$NE[C] = (1 - TP - FP) \times PCF(+) + FP \tag{5.25}$$

A point (TP, FP) representing a classifier in ROC space is converted by equation 5.25 into a line in cost space. A line in cost space, using equation 5.26 where S is the slope and TP_0 the intersection with the true positive rate axis. Both these operations are invertible. So there is also a mapping from points (lines) in cost space to lines (points) in ROC space. Therefore there is a bidirectional point line duality between the ROC and cost representations.

$$\begin{aligned} PCF(+) &= \frac{1}{1 + S}, \\ NE[C] &= (1 - TP_0)PCF(+). \end{aligned} \tag{5.26}$$

Figure 5.13 shows lines representing four extreme classifiers in the cost space. At the top is the worst classifier, it is always wrong and has a constant normalised expected cost of 1. At the bottom is the best classifier, it is always right and has a constant cost of 0.

The classifier that always chooses negative has zero cost when $PCF(+) = 0$ and a cost of 1 when $PCF(+) = 1$. The classifier that always chooses positive has cost of 1 when $PCF(+) = 0$ and a zero cost when $PCF(+) = 1$. Within this framework it is apparent that we should never use a classifier outside the shaded region of Figure 5.13 as a lower expected cost can be archived by using the majority classifier which chooses one or other of the trivial classifiers depending on PCF(+).

At the limits of the normal range of the probability cost function equation 5.25

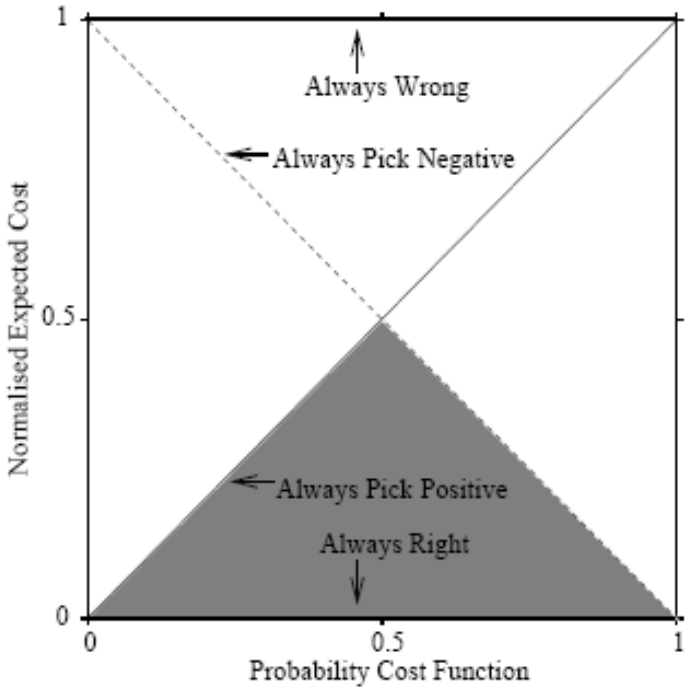

Figure 5.10: Cost space:classification

simplifies to $NE[C] = FP$ if $PCF(+) = 0$ and $NE[C] = (1 - TP)$ if $PCF(+) = 1$. To plot a classifier on the cost graph we set the point on the left hand side y-axis to FP and the point on the right hand side y-axis to $(1 - TP)$ and connect them by a straight line. Figure 5.14 shows a classifier with $FP = 0.09$ and $TP = 0.36$. The line represents the expected cost of the classifier over the full range of possible costs and class frequencies. This procedure can be repeated for a set of classifiers as shown in Figure 5.14. We can now compare the difference in expected cost between any two classifiers. There is no need for the calculations required in the ROC space, we can directly measure the vertical height difference at some particular probability cost value. Dominance is explicit in the cost space. If one classifier is lower in expected cost across the whole range of the probability cist function it dominates the other. Each classifier delimits a half space. The intersection of the half spaces of the set of classifiers gives the lower envelope indicated by the dashed line in Figure 5.15. This effectively chooses the classifier that has the minimum cost for a particular operating range. This is equivalent to the upper convex hull in the ROC space. This equivalence arises from the duality of the two representations.

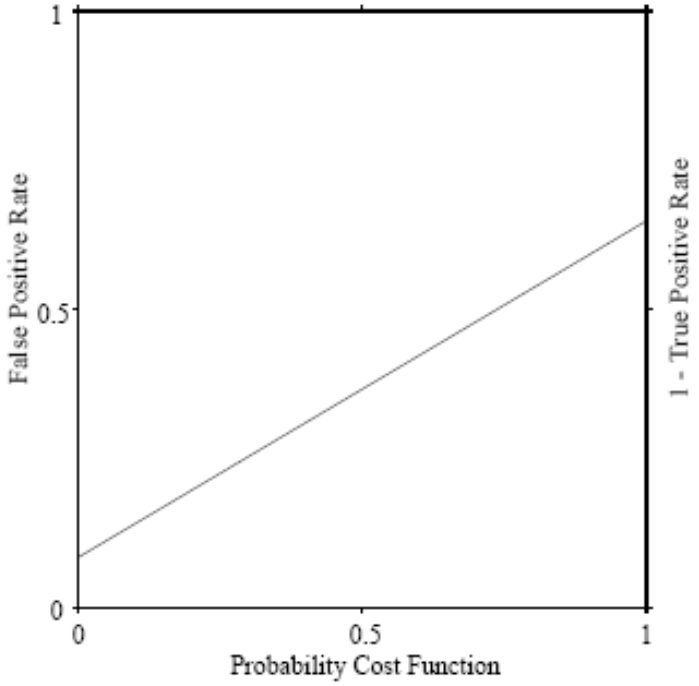

Figure 5.11: Cost space:comparison

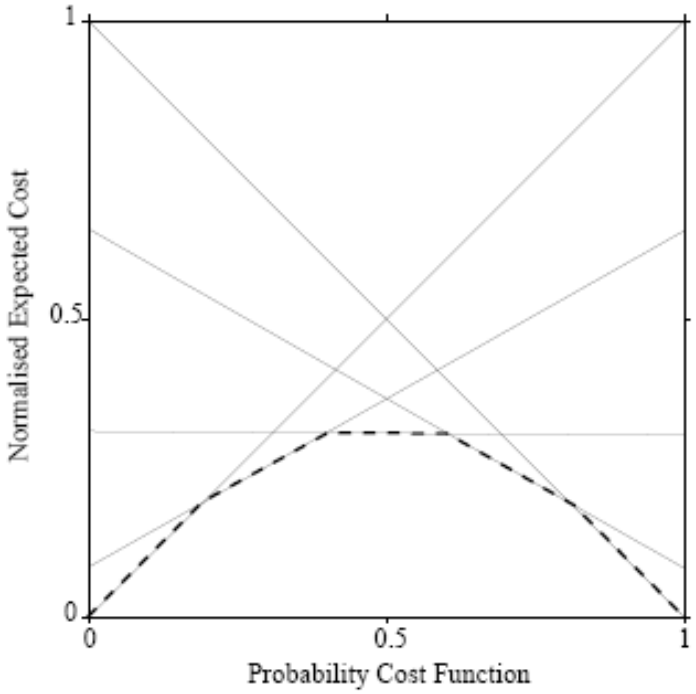

Figure 5.12: A set of classifiers

We remark that the literature presents other performance criteria based on error rate, area under the ROC curve, Neyman-Pearson criterion and workforce utilisation. For more details, see e.g. Provost et al. 1998. There are also many proposals on averaging multiple curves (see e.g. Swet and Picket 1982).

One interesting avenue of future research in this field is whether or not there are alternative dualities based on such metrics. The approaches proposed allows the easy calculation of the qualitative and quantitative model performance.

Chapter 6

Conclusions and Future research

In this Chapter we resume this dissertation. This research is a starting point that can be improved with new or old statistical methods. In particular, we report a review that can be applied to estimate customer lifetime value with the application of nonparametric techniques. This field of research can be studied and applied in a next time.

6.1 Conclusions

In the dissertation we presented a comparison between classical (static) and novel statistical techniques (dynamic) to predict rates of churn of customers.
The dissertation starts with a review on variable selection techniques. In Chapter 1 we review the literature on variable selection and we propose a general framework to improve the dimensionality reduction with attention on the relationships of Max-dependency, Max-relevance and Min-redundancy.
Chapter 2 starts with a mathematical formalisation to define Customer Lifetime Value. Later we introduce the classical statistical tenure model based on Cox regression and show the weaknesses and criticisms of such classical churn models. We then point attention to new lifetime value models, in particular we focus on survival analysis in the point processes framework. We put in this section also a review on model adequacy and model comparison.
In Chapter 3 we propose a Bayesian extension of lifetime value models. We propose a new Bayesian methods to choose the most important variables and we present a novel Bayesian model based on point processes framework (two step models). We

improve this model following some approach in literature based on Bayesian Model Averaging (one-step model). We then comment on the evolution of this approach, especially with stratification and multi-level models. We discuss model search and model comparison for Bayesian lifetime value models. We show the application of customer lifetime value models to a real marketing application.

Chapter 4 improves the previous results with the presentation of Bayesian stratified models, with fixed and random effects, and a discussion on the efficiency of partial likelihood methods. We then propose a new field of research (theoretical and applied) based on penalised likelihood methods, for which we propose a Bayesian extension and discuss on computational issues. We show that estimation using shared frailty models can be performed with penalized likelihood methods and therefore in a simple and realistic Bayesian framework. This is true for models with time-dependent covariates as well as for models with time-independent covariates on which we focused on in an attempt to keep the notation simple. We have found such a correspondence for gamma and gaussian frailty models; more research is needed for more general models. However we remark that the methodology support the use of AIC or corrected AIC as a selection criteria; with this approach, models can be fit beyond those for which a formal ML-penalized correspondence has been worked out, such as models with multiple frailty terms or other frailty distributions. Using AIC as the optimization criteria for θ and the log of a t-distribution density as the penalty term, for instance, appears to give similar results to more formal MCMC based on Bayesian methods on our application and formal results are needed to understand the relative merits of likelihood and degrees-of-freedom based approaches.

Beyond its extendibility, an important benefit of the penalized approach is computational speed. Because of the connection to other work in penalized regression, we have shown further computational improvements, possible for specific models. For example in shared frailty models, the use of a sparse Cholesky factorization provides significant computational advantages.

In Chapter 5 we propose new methods for model selection. The Chapter contains an idea of feature research. In particular we discuss on model assessment and predictive performance with particular attention to economic assessment and decision making, based on ROC Curves. We would like to investigate model priors when there is prior knowledge to be incorporated in the analysis and also averaging over model groups

and classes. In particular, we will consider uncertainty about functional model form, and uncertainty about the model for the baseline hazard function. It would be interesting to combine the methods for regression (variable search) and graphical models (structural search). A combination of these two problems would search over the space of variables and dependencies simultaneously, perhaps via techniques like Reversible Jump MCMC (Green, 1995), and could greatly extend the results presented here.

6.2 Future research

For the same problem, how to estimate customer lifetime value, we report here some new methods based on Bayesian non parametric inference that we like to study for next research.

In recent year, Bayesian nonparametric models, both theoretical and computational has witnessed considerable advances. The ear listed priors for nonparametric problems seem to have been described by Freedman who introduced tail-free and Dirichlet random measures. Subsequently Ferguson (1973,1974), Antoniak (1974), Diaconis and Freedman (1986), formalised and explored the notion of a Dirichlet process. Early work was largely focused on stylized summary estimates and tests so that comparisons with the corresponding frequentist procedures could be made.

An interesting approximation of the Dirichlet process is the procedure proposed by Muliere and Tardella (1998) and also another variation is the Dirichlet multinomial process introduced by Muliere and Secchi (1995).

More recently, the Beta-Stacy process was developed by Walker and Muliere (1997). Walker and Mallick (1996) detail the use of Polya trees, also in a frailty model (Clayton and Cuzick,1985). The Bayesian literature has grown rapidly. The focus of attention is on full Bayesian analyses of nonparametric models by using simulation techniques, apparently first used in this context by Escobar (1988). The paper of Walker, Damien, Laud and Smith (1999) discuss and illustrate the rich modelling and analytic possibilities that are available to the statistician within the Bayesian nonparametric and semiparametric framework. Other recent surveys of nonparametric Bayesian models appear in Dey, Muller and Sinha (1998). Nonparametric models based on Dirichlet process mixtures are reviewed in MacEachern and Muller

(1998). A recent review of nonparametric Bayesian inference in survival analysis can be found in Sinha and Dey (1997). We report in the next section a general review on Bayesian nonparametric survival models that can be employ in our next research to improve the estimation of Customer Lifetime Value.

Let $x_1, ..., x_n$ denote the survival times, $x_i \sim F(.)$. Let $C_1, ..., C_n$ denote the (possibly random) censoring times. The actually observed data is a collection of pairs $(T_1, I_1), ..., (T_n, I_n)$ with censored observations $T_i = min(x_i, C_i)$ and censoring indicator $I_i = (x_i \leq C_i)$. Two quantities are of primary interest in survival analysis: the survival function $S(t) = 1 - F(t)$ and the hazard rate function $\lambda(t) = F'(t)/S(t)$. In Bayesian nonparametric statistics there are more references of related approaches applying the Dirichlet process to similar problems see Ferguson, Phardia and Tiwari (1992). Doss (1994) studied an multivariate Dirichlet process for survival data subject to more general censoring schemes. Evaluation of the posterior mean on F is done through an interesting MCMC scheme that involves Dirichlet draws using a composition method.

Many stochastic process priors that have been proposed as nonparametric prior distributions for survival data analysis belong to the class of neutral to the right (NTTR) processes. There are many results in this framework. Doksum (1974) showed that the posterior for an NTTR prior and i.i.d. sampling is again an NTTR process. Ferguson and Phardia (1979) showed that for right censored data the class of NTTR process priors remains closed; that is the posterior is still an NTTR process. An alternative model was proposed by Hjort by placing a Beta process prior on $\Lambda(t) = \int_0^t \lambda(s)ds$. Full Bayesian inference for a model with a Beta process prior for the cumulative hazard function using Gibbs sampling can be found in Damien, Laud and Smith (1996). There are many references on this procedures.

A different modelling perspective is obtained by assuming dependence between hazards, for example with the introduction of a Markovian process prior on (λ_k), see Gamerman (1991).

To incorporates covariates, the most popular choice is the proportional hazards model, introduced in Cox (1972). Assuming $T_1, ..., T_n$ are the failure times of n individuals the hazard rate functions are modelled as:

$$\lambda_i(t) = \lambda_0(t)exp(Z_i(t)^T\beta), i = 1, ..., n, \tag{6.1}$$

where $Z_i(t)$ is the p-dimensional vector of covariates for the i-th individual at time $t > 0$, β is the vector of regression coefficient and $\lambda_0(t)$ is the baseline hazard rate function. A model based on an independent increments Gamma process was proposed by Kalbfleisch (1978) who studied its properties and estimation. In the context of multiple event time data, Sinha (1997) considered an extension of Kalbfleisch (1978). Accelerated failure time models are an alternative framework to introduce regression in survival analysis. The generic accelerated failure time model assumes that failure time T_i arise as $log(T_i) = -Z_i'\beta + log(x_i)$.

Nonparametric approaches assume a probability model for the unknown distribution of $log(x_i)$. Models based on Dirichlet process priors appear. Walker and Mallik (1999) propose an alternative prior model.

Bibliography

[1] Aalen O.O. and Hoem J.M. (1980) Random time changes for multivariate counting processes.*Scandinavian Actuarial Journal,* 81-101.

[2] Aalen, Odd O. (1989) A linear regression model for the analysis of life times. *Statistics in Medicine,* 8, 907-925.

[3] Amaldi E. and Kann V.(1998) On the approximation of minimizing non zero variabiles or unsatisfied relations in linear systems. *Theoretical Computers Science*, 209, 237-260.

[4] Anderberg M.R. (1973) *Cluster Analysis for Applications,* Academic Press, New York.

[5] Andersen P.K., Borgan O., Gill R.D. (1982) Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10, 1100-1120.

[6] Anderson J.A. and Blair V. (1982) Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika,* 69, 123-136.

[7] Anderson P. K., Borgan O., Gill R.D., and Keiding N. (1993) *Statistical Models Based on Counting Processes*. New York, Springer.

[8] Andersen P. K., and Gill R.D.(1982) Cox's regression model for counting processes: A large sample study. *The Annals of Statistics,* 10, 1100-1120.

[9] Antoniak C.E.(1974) Mixtures of Dirichlet processes with application to Bayesian nonparametric problems. *The Annals of Statistics,* 2, 1152-1174.

[10] Ball G.H. and Hall D.J. (1967) A clustering technique for summarizing multivariate data. *Behavl. Sci.,* 12, 153-155.

[11] Bauer, Daniel J. (2003) Estimating multilevel linear models as structural equation models.*Journal of Educational and Behavioral Statistics,* 28, 135-167.

[12] Beale E.M.L. (1969) Euclidean ckuster analysis. *Bull. I.S.I.,* 43, Book 2, 92-94.

[13] Bekkerman R., R. El-Yaniv, N. Tishby, and Y. Winter, (2003) Distributional word cluster vs. words for text categorization. *JMLR*, 3, 1183-1208.

[14] Berger, Paul D. and I. Nasr (1998) "Custumer Lifetime Value Marketing Models and Applications," *Journal of Interactive Marketing,* 97(1-2)12 (winter), 17-30.

[15] Berzuini C. and Clayton D.G. (1994) Bayesian analysis of survival on multiple time scales. *Statistics in Medicine,* 13, 823-838.

[16] Box G.E.P. (1980) Sampling and Bayes inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistic Society, Series A* , 143, 383-430.

[17] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) *Classification and Regression Trees*, Chapman and Hall.

[18] Breslow N.E. (1975) Analysis of survival data under the proportional hazards model. *International Statistical Review* , 43, 45-48.

[19] Brooks S.P. and Gelman A. (1998) General methods for monitoring convergence of iterative simulation. *Journal of Computastional and Graphical Statistical,* 7, 434-455.

[20] Brooks, S. P., Giudici, P. and Roberts, G. O. (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society, Series B*, 65, 3-39.

[21] Cai, J., Fan, J., Jiang, J. and Zhou, H. (2006) Partially linear Hazard regression for multivariate survival data. *to appear in Journal of American Statistical Association.*

[22] Cai, J., Fan, J, Zhou, H. and Zhou, Y. (2007) Marginal hazard models with varying-coefficients for multivariate failure time data. *to appear in The Annals of Statistics.*

[23] Calinski T. Hara J. (1974) A dendride method for cluster analysis. *Communications in Statistics*, 1-27.

[24] Carlin B.R and Chib S. (1995) Bayesian model choise via Markov Chain Monte Carlo methods. *Journal of Royal Statistical Society, Series B,* 57, 3, 473-484.

[25] Casella G., Moreno E. (2006) Objective bayesian variable selection. *Journal of the American Statistical Association* 101 pp. 157-167.

[26] Chen, K., Fan, J., Jin,Z. (2006) Design-adaptive Minimax Local Linear Regression for Longitudinal/Clustered Data. *to appear in Statistica Sinica.*

[27] Chen, J., Fan, J., Li, K.H. and Zhou, H. (2005) Local quasi-likelihood estimation with data missing at random. *to appear in Statistica Sinica.*

[28] Chen, Harrington and Ibrahim (2001) *Bayesian Survival Analysis,* Springer.

[29] Chipman H.A., George E.I. and McCulloch R.E. (1998) Bayesian CART model search. *Journal of the American Statistical Association* 93, 935-948.

[30] Clayton J. and Cuzick (1985) Multivariate generalizations of the proportional hazards model (with Discussion). *Journal of the Royal Statistical Society A* 148, 82-117.

[31] Clayton D. (1991) A Monte Carlo for Bayesian inference in frailty models. *Biometrics,* 47, 467-485.

[32] Clayton D. (1994) Bayesian analysis of frailty modes. *Technical Report, Medical Resource Council Biostatistics Unit*, Cambridge.

[33] Clifford H.T., and Stephenson W. (1975) *An Introduction to Numerical Classification,* Academic Press New York.

[34] Clyde M.A. (1999) Bayesian model averaging and model search strategies. In *Bayesian Statistics 6* (Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford University Press, pp. 157-185.

[35] Clyde M.A., (1999a) Discussion of Hoeting J.A., Madigan D., Raftery A.E., Volinsky C.T., Bayesian model averanging: a tutorial. *Statistical Science* 14, 409-412.

[36] Clyde M., DeSimone H. and Parmigiani G., (1996) Prediction via orthoganalized model mixing. *Journal of the American Statistical Association* 91 1197-1208.

[37] Clyde M.A. and George E.I., (1999) Empirical base estimation in walvelet nonparametric regression. In *Bayesian inference in Walvelet-Based Models* (P. Muller and B. Vidakovic, eds) 309-322. Springer.

[38] Clyde M.A. and George E.I., (2000) Flexible empirical Bayes estimation for wavelets. In *Journal of Royal Statistical Society, Series B,* 62, 681-698.

[39] Clyde, Merlise, Parmigiani, Giovanni and Vidakovic, Brani (1998) Multiple shrinkage and subset selection in wavelets. *Biometrika,* 85, 391-401.

[40] Constantine A.G., Gower J.C.(1978) Graphical representation of asymmetric matrices. *Applied Statistics,* 27, 297-304. 187-220.

[41] Cox D.R.(1972) Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34, 187-220.

[42] Cormack R.M.(1971) A review of classification. *Journal of the Royal Statistical Society, Series A* 134, 321-367.

[43] Cover T., Thomas J. (1991) *Elements of Information Theory.* Wiley.

[44] DeGroot M.H. (1970) *Optimal Statistical Decisions.* McGraw-Hill.

[45] Dey D., Muller P. and Sinha D., eds (1998) Practicle Nonparametric and Semiparametric Bayesian Statistics. *Lecture Notes in Statistic*, 133, Springer.

[46] Dhillon I. , Mallela S., and Kumar R. (2003) A divisive information-theoretic feature clustering algorithms for text classification *JMLR.*

[47] Diaconis P. and Freedman D. (1986) On the consistency of Bayes estimates. (with discussion). *The Annals of Statistics,* 14, 1-67.

[48] Doksum K. (1974) Tailfree and neutral random probabilities and their posterior distribution. *The Annals of Probabability,* 2, 183-201.

[49] Doss H. (1994) Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics,* 22, 1763-1786.

[50] Draper D. (1995) Assestment and propogation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B,* 57,45-70 .

[51] Duda R.O. ,Hart P. E., and Stork D.G. (2001) *Pattern Classification.* Wiley.

[52] Dunson (2005) Bayesian Semiparametric Isotonic Regression for Count Data *Journal of the American Statistical Association*, 100.

[53] Dunson, David B. and Herring, Amy H. (2005) Bayesian model selection and averaging in additive and proportional hazards models. *Lifetime Data Analysis,* 11, 213-232.

[54] Efron B. (1977) The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* 72, 557-565.

[55] Efron B. and Tibshirani R., (1993) *An Introduction to the Bootstrap,* Chapman and Hall.

[56] Efroymson M.A. (1960) Multiple regression analysis. *Mathematical Methods for Digital Computers,* Wiley.

[57] Escobar M. (1988) *Estimating the means of several normal population by estimating the distribution of the means.* Ph.D. dissertation, Dept. Statistics, Yale Univ.

[58] Everitt B.S. and Dunn G. (1991) *Applied Multivariate Data Analysis,* Edward Arnold, London.

[59] Fan J. and Li R. (2005) Variable selection via penalized likelihood, *Technical Report,* Department of Statistics, UCLA.

[60] Fan, J., Lin, H. and Zhou, Y. (2006) Local partial likelihood Estimation for life time data. *The Annals of Statistics*, 34, 290-325.

[61] Ferguson T.S. (1973) A Bayesian analysis of some nonpoarametric problems. *The Annals of Statistics* 1, 209-230.

[62] Ferguson T.S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics* 2, 615-629.

[63] Ferguson T.S. and Phadia E.G. (1979) Bayesian nonparametric estimation based on censored data. *The Annals of Statistics,* 7, 163-187.

[64] Ferguson T.S. and Phadia E.G. and Tiwari R.C. (1992) Bayesian nonparametric inference. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu,* (M. Ghosh and P.K. Pathak, eds.) 127-150. IMS, Hayward, CA.

[65] Fleming T.R. and Harrington D.P. (1991) *Counting Processes and Survival Analysis.* Wiley.

[66] Fogey E.W. (1965) Cluster analysis of multivariate data: efficiency versus interpretabily of classification . *Biometrics,* 21, 768-769.

[67] Forman G. (2003) Anextensive empirical study of feature selection metrics for classification. *JMLR*, 3, 1289-1306 .

[68] Foster D.P. and George E.I. (1994) The risk inflation criterion for multiple regression. *The Annals of Statistics,* 22, 1947-1975.

[69] Friedman H.P. and Rubin J. (1967) On some invariant criteria for grouping data. *Journal of the American Statistical Association,* 62, 1159-1178.

[70] Furnival G.M. and Wilson R.W. (1974) Regression by leaps and bounds. *Technometrics,* 16, 499-511.

[71] Gail M.H., Santner T.J. and Brown C.C. (1980) An Analysis of comparative cancirogenesis experiments with multiple times to tumor. *Biometrics,* 36, 255-266.

[72] Gamerman D. (1991) Dynamic Bayesan models for survival data. *Applied Statistics,* 40, 63-79.

[73] George E.I. (1999a) Discussion of Hoeting J.A., Madigan D., Raftery A.E., Volinsky C.T., Bayesian model averanging: a tutorial. *Statistical Science,* 4, 409-412.

[74] Gehan, Edmund A. and Thomas, Donald G. (1969) The performance of some two-sample tests in small samples with and without censoring. *Biometrika,* 56, 127-132.

[75] Gelman A. and Rubin D.(1992a) Inference from interative simulation using multiple sequences. *Statistical Science* 7, 457-511.

[76] Gelman A., GilksW.R., and Roberts G.O. (1996) Efficient metropolis jumping rules. In Berger J.O., Bernardo J.M, Dawid A.P., Lindley D.V., Smith A.F.M. (Eds.) *Bayesian Statistics, Vol. 5.* Oxford University Press, Oxford, pp. 599-608.

[77] Gelman A. and Rubin D.(1992b) A single from the Gibbs sampler provides a false sense of security. *Bayesian Statistics, Vol.4*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, New York: Oxford University Press, 625-631.

[78] George E.I. (1999) Discussion of "Model averanging and model search strategies" by M. Clyde. In *Bayesian Statistics, Vol. 6 (J.M. Bernardo, at al. eds),* 157-185. University Press, Oxford.

[79] George E.I. (2000) The variable selection problem. *Journal of the American Statistical Association,* 95, 452, 1304-1308.

[80] George E. I. and Foster D. P. (2000) Calibration and empirical Bayes variable selection. *Biometrika,* 87, 731-747.

[81] George E.I and McCulloch R.R. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association,* 88, 881-889.

[82] George E.I and McCulloch R.R. (1997) Approches for Bayesian variable selection. *Statististica Sinica,* 7, 339-373.

[83] George E., and McCulloch (1997) Variable selection via Gibbs sampling. *Journal of the American Statistical Association,* 88, 881-879.

[84] George, E. I. and McCulloch, R. E. (1999) Comment on "Variable selection and function estimation in additive nonparametric regression using a data-based prior". *Journal of the American Statistical Association,*94, 798-799.

[85] Geweke J. (1996) Bayesian reduced rank regression in econometrics. *Journal of econometrics,* 75, 121-146.

[86] Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," *in Bayesian Statistics,* eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Claredon Press, pp. 609620.

[87] Gower J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis.*Biometrika,* 53, 325-338.

[88] Gower J.C. (1985) Measures of similarity, dissimilarity and distance. In *Encyclopedia of Statistical science, Volume 5* (S.Kotz, N.L. Johnson and C.B. Read, eds).Wiley.

[89] Giudici P. (2003) *Applied data mining*, Wiley.

[90] Giudici P., Mezzetti M. and Muliere P.(2003) Mixtures of products of Dirichlet processes for variable selection in survival analysis *Journal of Statistical Planning and Inference,* 111, 101-115.

[91] Good I.J. (1952) Rational Decision. *Journal of the Royal Statistical Society, Series B,* 14, 107-114.

[92] Godsill, Simon J. (2001) On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics,*10, 230-248.

[93] Grambsch, Terry M. and Therneau, Patricia M. (2000) *Modeling survival data: extending the Cox model.*Springer.

[94] Grandvalet and S. Canu. (2002) Adaptive scaling for feature selection in SVMs. *Proceeding of NIPS 15.*

[95] Gray R.J. (1992) Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association,* 87, 942-951.

[96] Geen P.J.(1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. In *Biometrika,* 82, 711-732.

[97] Green P. and Silverman B., (1994) *Nonparametric Regression and Generalized Linear Models,* Chapman and Hall.

[98] Guti[U+FFFD]ez-Pe[U+FFFD]. (2005) Bayesian Methods for Categorical Data. *Encyclopedia of Statistics in Behavioral Science (B.S. Everitt y D.C. Howell, eds.)*, Vol. 1. Wiley.

[99] Harrington D.P., Fleming T.R. and Green S.J. (1991) Procedures for serial testing in censored survival data. In *Survival Analysis,* (Eds. J. Crowley and R.A. Jhonson). Hayward, CA: Institute of Mathematical Statistics, pp. 269-286.

[100] Hastie T.J., and Tibshirani R.J. (1990) Exploring the nature ovariate effects in the proportional hazard models. *Biometrics,* 46, 1005-1016.

[101] Hastie T., Tibshirani R., and Friedman J., (2001) *The elements of learning.*Springer.

[102] Hastie, Trevor, Tibshirani, Robert and Friedman, J. H. (2001) *The elements of statistical learning: data mining, inference, and prediction.*Springer.

[103] Hastings W.K. (1970) Monte Carlo sampling methods using Markow chains and their applications. *Biometrika,* 57, 1, 97-109.

[104] Hensler G. L., Mehrotra, K. G. and Michalek, J. E. (1977) A note on some exact efficiency calculations for survival distributions. *Biometrika,* 64, 635-637.

[105] Hocking, R. R. (1976) The analysis and selection of variables in linear regression, *Biometrics*, 32, 1-49.

[106] Hoerl A.E. and R.W. Kennard (1970) Ridge regression: biased estimation for nonortogonal problems. *Technometrics* 12, 55-67.

[107] Hoeting J.A., Madigan D., Raftery A.E. and Volisky C.T. (1999) Bayesian model averanging: a tutorial (with discussion). *Statistical science*, 14, 4, 382-417.

[108] Hoeting J.A., Madigan D., Raftery A.E. (1999a) Bayesian simoultaneous variable and trasformation selection in linear regression. *Thecnical report 9905,* Department of Statistics, Colorado State University.

[109] Hoeting J.A., Madigan D., Raftery A.E. (1996) A method for simultaneous variable selection and outlier identification in linear regression.*Computastional Statistics and Data Analysis,* 22, 251-270.

[110] Hougaard P.(1995) Frailty model for survival data. *Lifetime Data Analysis*, 1, 255-273.

[111] Hougaard P. (2000) *Analysis of Multivariate Survival Data.*Springer.

[112] Hosmer, David W. and Lemeshow, Stanley (1999) *Applied survival analysis: regression modeling of time to event data*. Wiley.

[113] Huang, Jie and Harrington, David (2002) Penalized partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter. *Biometrics,* 58, 781-791.

[114] Ibrahim J.G, Chen M-H., and Sinha D. (2001a) Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*.

[115] Ibrahim J.G., Chen M.-H., anfd Sinha D. (2001b) Criterion based methods for Bayesian model assessment. *Statistica Sinica*.

[116] Jancey R.C. (1966) Multidimensional group analysis.In *Anust. J. Bot.,* 14, 127-130.

[117] Jebara T. and Jakkola T.(2000) Feature selection and dualities in maximum entropy discrimination. In *Annual Conference on Uncertainty in artificial intelligence*.

[118] Jones,M. P. (1996) Indicator and Strati. cation Methods for Missing Explanatory Variables in Multiple Linear Regression, *Journal of the American Statistical Association*, 91, 222-230.

[119] Kalbfleisch J.D. (1978) Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* 40, 214-221.

[120] Kalbfleisch J.D., and Prentice R.L. (1980) *The statistical analysis of failure time data.* Wiley.

[121] Kaplan E.L. and Meier P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* 53, 457-481.

[122] Karrison, Theodore (1987) Restricted mean life with adjustment for covariates *Journal of the American Statistical Association,* 82, 1169-1176 Keywords: Survival distribution; Censored data; Cox regression model.

[123] Kass R.E. and Raftery A.E., (1995) Bayes factor. *Journal of the American Statistical Association,* 90, 773-795.

[124] Kass R.E. and Wasserman L., (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association,* 90, 928-934.

[125] Key, J.T, pericchi, L.R. and Smith A.F.M. (1999) Bayesian model choise: what and why? (with discussion) In *Bayesian Statistics vol.6,* 343-370.

[126] Klein J.P. and Moeschberger M.L. (1997) *Survival Analysis* Springer.

[127] R. Kohavi and G. John (1997) Wrappers for feature selection. *Artificial intelligence*, 97(1-2), 273-324, December .

[128] Kuk A.Y.C.(1984) All subsets regression in aproporzional hazards model. *Biometrika* 71, 587-592.

[129] Krzanowski W.J. (1988) *Principal of Multivariate Analysis: A User's Perspective,* Oxford University Press, Oxford.

[130] Laud P.W. and Ibrahim J.G. (1995) Predictive model selection. *Journal of the Royal Statistical Society, Series B* 57, 247-262.

[131] Laud P.W., Smith A.F.M., and Damien P. (1996) Monte Carlo methods for approximating a posterior hazard rate process. *Statistics and Computing,* 6, 77-83.

[132] Laud P.W and Ibrahim J.G. (1996) Predictive specification of prior model probabilities in variable selection. *Biometrika,* 83-2, 267-274.

[133] Lawless J.F. and Singhal K. (1978) Efficient screening of nonnormal regression models. *Biometrika* 34, 318-327. Springer.

[134] Leamer E.E,. (1978) *Specification Searches* Wiley.

[135] Lee E.W., Wei L.J. and Amato D.A., (1992) *Cox-type regression analysis for large numbers of small groups of correlated failure time observations.* In Suirvival Analysis: State of the Art, Ed. J.P. Klein and P. Goel, pp. 237-48. Wiley.

[136] Lewis S.M. and Raftery A.E. (1997) Estimating Bayes factors via posteriors simulation with the Lapace-Metropolis estimator. *Journal of the American Statistical Association,* 92, 648-655.

[137] MacEachern S.N. and Muller (1998) Estimating mixture of Dirich process model. *Journal of computional and graphical statistics,* 7,223-238.

[138] MacQueen J. (1967) Some methods for calssification and analysis of multivariate observations. *Proceeding 5th Berkeley Symp.,* 1, 281-297.

[139] Madigan D. and Raftery A.E.(1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association,* 89, 1535-1546.

[140] Madigan D. and York J. (1995) Bayesian graphical models for descrete data. *International Statistical Review,* 63, 215-232.

[141] Madigan D. and Raftery A.E. (1994) Model selection and accounting formodel uncertainty in graphical models using Occams window.*Journal of the American Statistical Association,* 89, 1535-1546.

[142] Mallows C.L. (1973) Some comments on C_p *Technometrics,* 15, 661-676.

[143] May, Susanne and Hosmer, David W. (1998) A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis,* 4, 109-120.

[144] MacQueen J. (1967) Some methods for classification and analysis multivariate observations *Proceeding 5th Berkeley Symp.,* 1, 281-297.

[145] Marquardt D. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics,* 11, 431-441, 1963.

[146] Marriot F.H.C. (1971) Practical problems in a method of cluster analysis. *Biometrics,* 27, 501-514.

[147] Marriot F.H.C. (1982) Optimization methods of cluster analysis. *Biometrika,* 69, 417-421.

[148] McGilchrist C.A. (1983) REML estimation for survival models with frailty. *Biometrics* 49, 221-225.

[149] McQuarrie A.D.R. and Tsai C.L. (1998) *Regression and Times Series. Model Selection.* World Scientific, Singapore.

[150] Meyer D.R. and Wilkinson R.G. (1998) Bayesian variable assessment.*Communications in statistics - Theory and Methods.* 27, 11, 2675-2705.

[151] Miller A.J. (1990) *Subset Selection in Regression,* Chapman and Hall.

[152] Milligan G.W. and Cooper M.C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika,* 50, 159-179.

[153] Mitchell T.J and Beauchamp J.J. (1988) Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* 91, 142-153.

[154] Muliere P. and Secchi P. (1995) A note on a proper Bayesian bootstrap. Technical Report 18, Dipt. Economia Politica e Metodi Quantitativi, Univ. Studi di Pavia.

[155] Muliere P. and Tardella L. (1998) Approximating distribution of random functionales of Fergusson-Dirichlet priors. *Canadian Journal of Statistics.* 26, 283-297.

[156] Muliere P. ans Walker S.G. (1997) A Bayesian nonparametric approach to survival analysis using P`[U+FFFD]` trees. *Scandinavian Journal of Statistics.* 24, 331-340.

[157] Nielsen G.G, Gill R.D., Andersen P.K. and Sorensen (1992) Accounting process approach to maximum likelihood estimation failty models. *Scandinavian Journal of Statistics.* 19, 25-43.

[158] Noble R. (2000) *Multivariate Applications of Bayesian Model Averanging.* PhD Dissertation, 2000, Virginia Polytecnic Institute.

[159] Oakes D.(1977) The asymptotic information in censored survival data. *Biometrica* 64, 441-448.

[160] Oakes D.(1992) Frailty models for multiple event times. In *Survival Analysis: State of the Art* (Eds J.P. klein and P.K. Goel). Netherlands: Kluwer Academic, pp.371-379.

[161] Parzen, Michael and Lipsitz, Stuart R. (1999) A global goodness-of-fit statistic for Cox regression models. *Biometrics,* 55, 580-584.

[162] Peace, Karl E. and Flora, Roger E. (1978) Size and power assessments of tests of hypotheses on survival parameters. *Journal of the American Statistical Association,* 73, 129-132.

[163] Pe[U+FFFD]uttman, Irwin, Daniel and Redondas, Dolores (2005) A Bayesian approach for predicting with polynomial regression of unknown degree.*Technometrics,* 47, 23-33.

[164] F. Pereira, N. Thishby, and L.Lee. (1993) Distributional clustering of English words. In *Proceeding Meeting of the association for Computional Linguistics,* pages 183-190.

[165] S. Perkins, K. Lacker, and J. Theiler.(2003) Grafting: Fast incremantal feature selection by gardient descend in function space. *JMLR*, 3, 1333-1356.

[166] Pfeifer, Christian (2004) Classification of longitudinal profiles based on semiparametric regression with mixed effects *Statistical Modelling,* 4, 314-323

[167] Foster Provost and Tom Fawcett, (1998) Robust classification systems for imprecise environments, *in Proceedings of the Fifteenth National Conference on Artificial Intelligence: in Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 706 713, Menlo Park, CA.

[168] Foster Provost, Tom Fawcett, and Ron Kohavi, (1998) The case against accuracy estimation for comparing induction algorithms, *in Proceedings of the Fifteenth International Conference on Machine Learning,* pp. 43 48, San Francisco. Morgan Kaufmann.

[169] Raftery A.E. (1995) Bayesian model selection in social research (with discussion). In *Sociological Methodology.* Cambridge, Massachussetts: Blackwells, pp 111-195.

[170] Raftery A.E. (1996) Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika,* 83, 251-266.

[171] Raftery A.E., Madigan D. and Hoeting J.A. (1997) Bayesian model averanging for lineal regression models. *Journal of the American Statistical Association,* 92, 179-191.

[172] Raftery, Adrian E. (1999) Comment on "A critique of the Bayesian information criterion for model selection". *Sociological Methods and Research,* 27, 411-427.

[173] Raftery A.E., Madigan D., and Volinsky C.T. (1995) Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics vol. 5* (Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith).Oxford University Press, pp 323-350.

[174] Rao C.R. (1952) *Advanced Statistical Methods in Biometrics Research,*Wiley.

[175] Reichheld, frederick f. (1996) *The Loyalty Effect: The Hidden Force Behind Growth, Profits and Lasting Value,* HBS Press, Boston.

[176] Rondeau, Virginie, Commenges, Daniel and Joly, Pierre (2003) Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis,* 9, 139-153.

[177] Rudemo, Mats (1982) Empirical choice of histograms and kernel density estimators.*Scandinavian Journal of Statistics,* 9, 65-78.

[178] Sahu S.K., Dey D.K. Aslanidou H. and Sinha D. (1997) A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis,* 3, 123-137.

[179] Schwarz G. (1978) Estimating the dimension of a model. *The Annals of Statistics* 6, 461-464.

[180] Shao J. (1996) Bootstrap model selection. *Journal of the American Statistical Association.* 91, 434, 655-665.

[181] Shively T.S., Kohn R. and Wood S. (1999) Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussion). *Journal of the American Statistical Association* 94, 777-806.

[182] Silverman B.W. (1985) Some aspects of the spline smoothing approch to non-parametrich regression curve fitting (with Discussion). *Journal of the Royal Statistical Society, Series B* 47, 1-52.

[183] Singleton R.C. and Kautz W. 1965 *Minimum squared error clustering algorithm.* Stanford Research Institute.

[184] Sinha D., Chen M-H, and Ghosh S.K. 1999 Bayesian analysis and model selection for interval-censored survival data. *Biometrics 55*, 585-590.

[185] Sinha D., Debajyoti, Ibrahim, Joseph G. and Chen, Ming-Hui (2003) A Bayesian justification of Cox's partial likelihood *Biometrika,* 90, 629-641.

[186] Sinha D., and Dey D.K. (1997) Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association,* 92, 1195-1212.

[187] Smith A.F.M, kohn R. (1996) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* 75, 317-343.

[188] Smith A.F.M, and Roberts G.O. (1993) Bayesian computation via the Gibbs sampler and related Markow chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B,* 55, 1, 3-23.

[189] Smith A.F.M, Spiegelhalterd J. (1980) Bayes factors as choise criteria for lineal models. *Journal of the Royal Statistical Society, Series B,* 42, 2, 213-220.

[190] Sneath P.H.A. and Sokal R.R. (1973) *Numerical Taxonomy,* W.HFreeman and Co., San Francisco.

[191] Swets J.A. (1988) Measuring the accuracy diagnostic system. *Science* 240, 1283-1293.

[192] John A. Swets and Ronald M. Pickett (1982), *Evaluation of diagnostic systems : methods from signal detection theory,* Academic Press, New York.

[193] Spath H. (1985) *Cluster Dissection and Analysis,* Ellis Horwood Ltd., chichester.

[194] Tanner M.A. and Wong W.H. (1987) The calculation of posterior distributions data augmentation (with discussion). *Journal of American Statistical Association,* 82, 528-550.

[195] Therneau T.M., Grambsch P.M., and FlemingT.R. (2000) Martingale based residuals for survival models. *Biometrika* 77, 147-160.

[196] N. Tishby, F. C. Pereira, and W. Bialek. (1999) The information bottleneck method. I *Proceeding of the 37th Annual Allerton Conference on communication, control and Computing,* pages 368-377.

[197] Tibshirani R.(1997) Regression shrinkage. *Journal of the royal statistical society, Series B*, 58, 267-288.

[198] Torkkola K. ,(2003) Feature extraction by non-parametric mutual information maximization. *JMLR*, 3, 1415-1438.

[199] Vannucci M., Brown P. J., and Fearn T. (1998a). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B,* 60, 627-641. (1998b) Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics,* 12, 173-182.

[200] Vannucci M., Brown P. J. and Fearn, T. (2001) Predictor selection for model averaging. *In Bayesian Methods with Applications to Science, Policy and Official Statistics* (eds E. I. George and P. Nanopoulos), pp. 553 562.

[201] Van Rijsbergen C. J., (1979) *Information retrieval,* London.

[202] Vaupel J.M., Manton K.G. and Stallard E. (1979) The impact of heterogeneity in individual frailty on the dynamics of morality. *Demography* 16, 439-454.

[203] Brown, P. J., Vannucci, M. and Fearn, T. (1998a) Multivariate Bayesian variable selection and prediction. J. R. Statist. Soc. B, 60, 627-641. (1998b) Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, 12, 173-182.

[204] Verweij, Pierre J. M. and Van Houwelingen, Hans C. (1993) Cross-validation in survival analysis.*Statistics in Medicine,* 12, 2305-2314.

[205] Volinsky C.T., Madigan D., Raftery A.E., and Kronmal R.A. (1997) Bayesian model averanging in proportional hazard modes: Assessing the risk of a stroke. *Applied Statistics* 46, 433-448.

[206] Walker S.G. and Mallick B.K. (1996) Hierarchical generalised linear models and frailty model with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B* 59, 845-860.

[207] Walker S.G. and Damien. (1998) A full Bayesian nonparametric analysis involving a neutral to the righ process. *Scandinavian Journal of Statistics,*25, 669-680.

[208] Walker S.G. and Damien P. and Laud P. and Smith A.F.M. (1999) Bayesian nonparametric inference for random distributions and relative functions (with discussion) . *Journal of the Royal Statistical Society, Series B,* 61, 485-527.

[209] Walker, S.G. y Guti[U+FFFD]ez-Pe[U+FFFD]. (2006) Bayesian Parametric Inference in a Nonparametric Framework. *to appear in Test*

[210] Walker S.G. and Mallick B.K. (1997) Hierarchical generalised linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B* 59, 845-860.

[211] Walker S.G. and Muliere P. (1997a) A characterization of P[U+FFFD] tree distibution. *Statistics and Probabily Letters,* 31, 163-168.

[212] Wallace C.S. and Boulton D.M. (1968) An information measure for classification. *Computer Journal,* 11, 185-194.

[213] Wei L.J., Lin D.Y. and Weissfeld L. (1989) Regressionan alysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association,* 84, 1065-73.

[214] Weston J., Elisseff A. , Schoelkopf B., and Tippinuse M. (2003) Of the zero norm with linear models and kernel methods. *JMLR*, 3, 1439-1461.

[215] Weston J. , Mukherjee S., Chapelle O., Pontil M., Poggio T., and Vapnik V. (2000) Feature selection for SVMs. In *Proceedings of NIPS*.